格致方法
定量研究系列

· 革新研究理念
· 丰富研究工具
· 最权威、最前沿的定量研究方法指南

吴晓刚 / 主编
【美】乔治 · H. 邓特曼(George H. Dunteman) / 等著

广义线性模型

格致出版社 上海人民出版社

世纪出版

格致方法
定量研究系列

广义线性模型

《广义线性模型》由四本介绍线性模型的小册子组成，它们分别是《广义线性模型导论》、《应用logistic回归分析》、《定序因变量的logistic回归模型》以及《logit 与probit：次序模型和多类别模型》。全书集中介绍了社会学研究分析方法中的一个非常有效且重要的数据分析方法，即线性模型。作者通过阐释广义线性模型的概念、基本原则，探讨了从线性模型推广至其他模型的可能路径，并举例比较了不同模型的拟合优度，为读者全面掌握线性模型分析法提供了一个可行的指南。

上架建议：社科研究方法
ISBN 978-7-5432-1901-4
9 787543 219014 >
定价：59.00元
易文网：www.ewen.cc
格致网：www.hibooks.cn

格致方法
定量研究系列

广义线性模型

格致出版社 上海人民出版社

图书在版编目(CIP)数据

广义线性模型/(美)邓特曼(George H. Dunteman)等著；吴晓刚主编. 一上海：格致出版社：上海人民出版社，2011
(格致方法·定量研究系列)
ISBN 978-7-5432-1901-4

Ⅰ. ①广… Ⅱ. ①邓… ②吴… Ⅲ. ①线性模型-研究 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2010)第 254752 号

责任编辑 高 璇
封面装帧 人马艺术工作室·储平

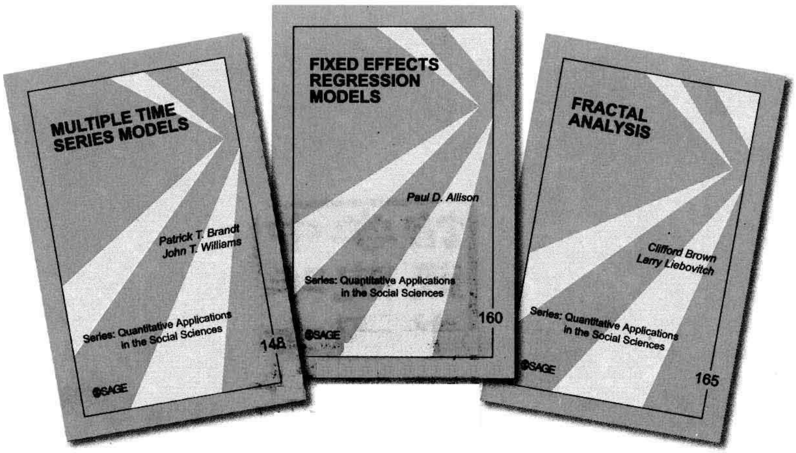
格致方法·定量研究系列
广义线性模型
[美]乔治·H. 邓特曼 等著
吴晓刚 主编

出 版 世纪出版集团 www.ewen.cc
格致出版社 www.hibooks.cn
上海人民出版社
(200001 上海福建中路193号24层)

编辑部热线 021-63914988
市场部热线 021-63914081

发 行 世纪出版集团发行中心
印 刷 上海图宇印刷有限公司
开 本 787×1092毫米 1/16
印 张 24.5
插 页 1
字 数 365,000
版 次 2011年8月第1版
印 次 2011年8月第1次印刷
ISBN 978-7-5432-1901-4/C·46
定 价 59.00元

绿皮书的传奇



1976 年，当政府管理与政治学教授 Eric Uslaner 向 SAGE 出版社的创始人 Sara 建议出版关于重要定量研究方法的简明小册子时，没有人预见到这套书会取得巨大成功。

那年夏天，“社会科学定量研究方法” 丛书诞生。统一的、朴素的绿色封面，上面仅仅印着书名、作者名及本书所属的系列名。Iverson 和 Northporth 的《变量分析》(*Analysis of Variance*)、Nagel 的《运筹学》(*Operations Research*)以及 Henkel 的《显著性检验》(*Tests of Significance*)是最早出版的几本，每本售价 2.95 英镑，被形容为 “像烤饼一样好卖”。

SAGE 选择了 20 种顶尖的定量研究工具，然后去寻找合适的作者，邀请他们围绕这些工具撰写 92 页的小书。这些薄薄的绿皮书在课堂上深受欢迎，在图书馆成套陈列。

至今，绿皮书系列中共有 160 种在初版或再版。书的主题反映了量化研究方法的发展：从基础统计知识、数据类型、测量到计算机的应用以及博弈论。这套书非常畅销，其中最畅销的一本，是 1980 年出版的 Michael Lewis-Beck 的《回归方法的应用》(*Applied Regression*)。

出版说明

《广义线性模型》是"格致方法 · 定量研究系列"丛书中的一本，由四本介绍线性模型的小册子构成，它们分别是《广义线性模型导论》、《应用 logistic 回归分析》、《定序因变量的 logistic 回归模型》以及《logit 与 probit：次序模型和多类别模型》。

《广义线性模型导论》系统地介绍了广义线性模型的概念、基本原则等内容，并探讨了如何从线性回归模型推进到非连续自变量的其他模型，又不失模型间的共同根基和相似性。《应用logistic回归分析》全面解释了 logistic 回归模型的估计、解释和诊断结果，并比较了估计标准误差和斜率等概念。作者运用了最新的统计软件，令人信服地评论了不同模型的拟合优度。最后，作者还详细说明了多选项和不排序多分类因变量的问题。《定序因变量的 logistic 回归模型》以展示来自儿童早期追踪研究的实例开始，通过呈现三种形式的因变量集中分析 logit 模型。随后，作者进而回顾了 logistic 回归，并列举了 SAS 和 SPSS 的案例。《logit 与 probit：次序模型和多类别模型》致力于分析当因变量具有多种类别时的估计情况。继第 1 章概论后，作者集中关注了具离散和次序形式的因变量，并比较了不同方法间的差异和优劣。最后，作者把处理对象扩展到具有两个以上结果的多类别或非次序因变量，为估计和解释从更复杂的离散因变量模型中得出结果提供了一个可行的指南。

总 序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996 年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫"调查数据分析"，第三门叫"研究设计"。除此以外，还有如"定类数据分析"、"人口学方法与技术"、"事件史分析"、"多层线性模型"等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。"社会科学里的统计学"(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。"社会科学中的定量分析"则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的"定量社会科学研究"丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选

一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

由于所选每本书都有一篇序言，对相关方法的背景和应用作了很好的介绍，我们均予以保留，内容在此不再赘述。为了方便起见，我们将内容相似的书目集册出版，每册三至五本不等，共八册，它们分别是：《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容，读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版，能为推动国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

目 录

广义线性模型导论

序 3
第 1 章　广义线性模型 4
第 2 章　一些基础的模型化概念 9
第 3 章　经典多元回归模型 14
第 4 章　广义线性模型的基本原则 22
第 5 章　最大似然估计 28
第 6 章　离差和拟合优度 36
第 7 章　logistic 回归 41
第 8 章　泊松回归 50
第 9 章　生存分析 59
第 10 章　结论 68
附录 69
参考文献 77
译名对照表 78

应用 logistic 回归分析

序 81

第 1 章　线性回归和应用 logistic 回归模型 82
第 2 章　评估 logistic 回归模型的统计概要 96
第 3 章　解释 logistic 回归系数 116
第 4 章　诊断 logistic 回归的介绍 136
第 5 章　多分类 logistic 回归及其替代方法 155
附录　概率 164
注释 165
参考文献 167
译名对照表 171

定序因变量的 logistic 回归模型

序 177
第 1 章　概述 179
第 2 章　背景介绍：儿童早期追踪研究 184
第 3 章　背景：logistic 回归 189
第 4 章　定序结果的累积(比例)比数模型 206
第 5 章　连续比例模型 233
第 6 章　相邻类别模型 255

第 7 章　结论 264
附录一　第 3 章 268
附录二　第 4 章 270
附录三　第 5 章 272
附录四　第 6 章 276
注释 278
参考文献 279
译名对照表 284

logit 与 probit：次序模型和多类别模型

序 289
第 1 章　概论 291
第 2 章　次序模型 295
第 3 章　多类别模型 332
第 4 章　STATA 程序列表 358
注释 374
参考文献 377
译名对照表 379

广义线性模型导论

作者简介

乔治·H. 邓特曼(George H. Dunteman)

曾为美国三角研究院(Research Triangle Institute)首席科学家,致力于社会和行为科学领域的应用研究。曾在教育考试服务(Educational Testing Service)及美国军人研究机构(U. S. Army Research Institute)担任研究职务,并曾在罗切斯特大学担任助理教授,在佛罗里达大学担任副教授。1987—1988 年,担任维克森林大学 Babcock 管理学院的管理学访问教授,并教授 MBA 量化方法的核心课程。此外他还曾担任《教育及心理学测量》(*Educational and Psychological Measurement*)杂志的编辑并在专业期刊中发表多篇论文。

何满镐(Moon-Ho R. Ho)

加拿大麦吉尔大学心理学助理教授,并在新加坡南洋理工大学人文及社会科学学院、新加坡国家教育机构教学法及实习研究中心担任助理教授,教授本科生和研究生程度的进阶统计学课程。其研究兴趣包括:对于评断及决策决定的认知模型、态度之形成、构成健康相关行为改变之机制以及父亲参与儿童教养等。

译者简介

林毓玲

曾先后在台湾大学获本科学位、在美国芝加哥大学获硕士学位,现为香港科技大学硕士研究生。

序

本书编辑过程并不寻常:作者及编者都有所改变。我的前一任编辑,迈克尔·刘易斯·贝克,非常睿智地看到《广义线性模型导论》的价值。在2004年初从主编岗位退下来前,他看遍了计划书及先前的手稿版本。令人难过的是,乔治·H.邓特曼在完成他所认为的终稿后就过世了。进一步的修改由何满镐接手,他非常勇敢地接受挑战,并对原稿作出许多重要的修正。

社会科学家所分析的结果变量可以是连续的或是离散的。在已出版的丛书中,有许多书目涉及需要处理一个连续的因变量(及一些重要假设)的模型,经典线性回归为这类模型的代表。除此之外也涉及因变量是非连续的,通常统计模型的对象为事件发生几率,但也可能是频率或是对数频率。在过去20年中,许多型态的logit、probit(及对数线性)模型已经成为社会科学家众多分析方法中的标准,并且该丛书中也有多本涉及这些主题。

连续结果变量及离散的因变量这两种模型间的关系,在广义线性模型的架构下变得清晰。在社会科学中,研究者对在方程右方以 x 和 β 线性组合所表示的可线性化的自变量比较熟悉。然而,位于这两种模型左方的因变量 y 可以是多种形式的,包括metric、二元的、序列的、multinomial和计数的。再者,两种模型中的随机结果 y 可能服从正态、二项、泊松、gamma分布,且所有这些分布都属于指数家族分布。一旦我们对于 y 的随机分布作出服从指数分布的适当假设后,剩余的任务便是指明随机变量的期望以及 x 和 β 线性组合间的关系。将期望的随机结果变量对应到 x 和 β 的线性组合,是广义线性模型的一部分。

本书的根本目标是:对于熟悉经典线性回归的普通社会科学研究者,要如何从线性回归模型推广到非连续自变量的其他模型,而不失两种模型间的共同根基及相似性?本书两位作者陪着读者走访这一过程,并在沿途中启蒙不识此道者,这也对丛书提供了有益的增补。

——廖福挺(Tim Futing Liao)

第 1 章 | 广义线性模型

广义线性模型，顾名思义，为经典线性回归模型的普遍化。经典线性回归模型假设因变量为一组自变量的线性方程，且因变量为连续且正态分布的，有固定的方差。自变量则可以是连续的、类别的或两者的组合。多元回归分析、方差分析及协方差分析皆为线性模型的经典例子。它们皆可被写成：

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

其中 y 是连续性因变量，X_j 是自变量，ε 为假设正态分布的误差。因变量由两部分组成：系统性(systematic)成分 $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$；误差成分 ε。系统性成分即在任意组给定的 X_j 的值之下，y 的期望值 $E(y)$。即：

$$E(y \mid X_1, \cdots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

它是一个给定 X_j 值的条件下的平均数。回归分析的目的就是寻找一组以拟合优度来衡量具有高度解释力的自变项，即我们能凭借自变量的线性组合来解释 y 大部分的变异。如果回归参数 β_j 很大，当 X_j 的值从一观察值变化至另一观察值时，y 的期望值或 y 的条件平均数 (conditional mean)也将有很大的变异。如果在条件平均数或预测值中的变异比在 ε 中的变异更大，我们则能利用一个有用的模型，在给定自变量取值的条件下预测 y 的取值，以及了解不同自变量在解释因变量 y 的变异时的相对重要性。图 1.1 呈现了一个简单的线性回归模型($\beta_0 = 1$, $\beta_1 = 1.5$)。我们通过观察对象的一个随机样本，收集 y 的测量值以及 X_1, X_2, $\cdots$, X_p 来估计回归参数 β_j。就观察目的而言，我们的观察对象通常是人，但在其他应用中，观察对象可以是任何事物，如树、牛，甚至河流。

如果我们以 i 标示人，以 j 标示变量，则可以通过最小化误差的平方和来估算 β_j。

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ij})^2$$

在此，小标 i 被用以强调自变量的值随着个体（subject）的不同而变化的事实。此回归参数的估计方法通常被称做普通最小二乘法。

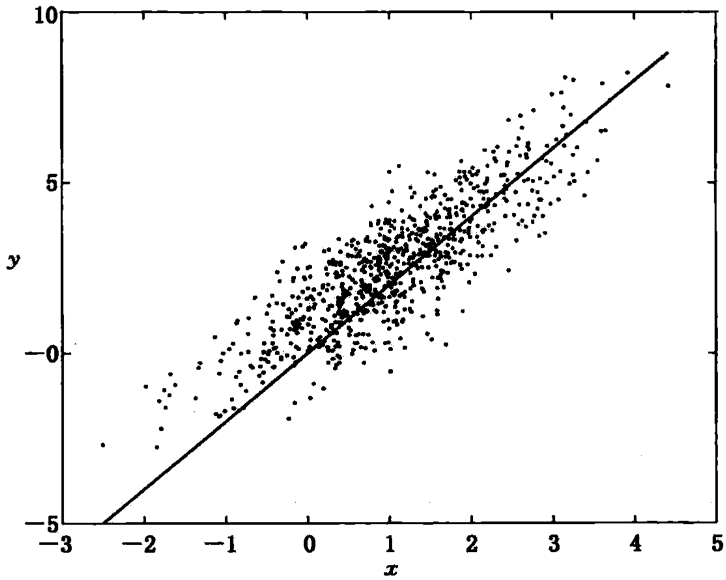


图 1.1 线性回归模型

这个线性回归模型自 19 世纪初步发展以来，对社会科学及其他科学特别有用。它很易于公式化，易于理解，并且回归系数易于利用普通最小二乘法估计。因此，它至今仍被广泛应用于各学科中。虽然它假设误差是正态分布的，但当误差接近于正态分布时，它仍是稳健的。

然而，在过去几十年中，人们已经广泛地意识到线性回归模型的局限。它假设因变量为连续或至少是准连续的，如考试成绩、个人特质测量等。而且，它假设该连续变量至少是接近于正态分布，并且其方差并不是其平均数的函数。内尔得和韦德伯恩（Nelder & Wedderburn, 1972）提出了广义线性模型，后来发

展为应用于非正态因变量的回归模型。

在许多应用中，因变量是类别的或包含计数的，抑或为连续的但并非正态。一个类别的因变量的例子是二元变量，只有两个离散的值 0 或 1，其中，1 代表事件的发生（如从大学中退学），而 0 代表事件未发生（如未从大学中退学）。目标是要模型化感兴趣的事件的发生概率。在稍后会提及 logistic 回归，它是广义线性模型的一种，适合此类型的数据。

一个关于计数的因变量的例子是，一个药物滥用者群体在五年里的药物滥用事件（treatment episodes）。我们将再一次地展示泊松回归（poisson regression），这是适合此情形的另一种广义线性模型。在这两个例子中，因变量都不是连续的，更不是正态分布的，且 0—1 二元变量与计数变量都为非负数。然而，在一般回归中连续因变量可以是正值或负值。

一个被广泛应用的非正态连续分布的例子为 gamma 分布。gamma 分布是偏斜的（skewed），只有正值，且其方差为其平均数的函数。它可以用来模型化一般性的、类别的、只有正值的因变量，如收入、生存时间及雨量。因变量为 gamma 分布的模型可以被置于广义线性模型的架构之下。

要注意，对于一个给定分布的因变量而言，自变量可以有许多种分布形式，且它们并不需要和因变量具有相同的分布。例如，与一正态分布因变量相关联的自变量可以包含许多不同的非正态分布，如均匀分布和多峰分布。如前所述，一般回归假设 y 的平均值会随着自变量变化，但关于条件平均数的 ε 的变异则维持不变。对于二元变量和计数变量来说，条件平均数的方差为其平均数的函数。如二元变量，因变量条件平均数为概率 p（例如，事件 1 发生的概率），而此平均数的方差为 $p(1-p)$，是平均数 p 的一个函数。因为平均数 p 会作为自变量的一个方程而变化，此二元变量的方差也会如此。对计数变量而言，泊松分布常被使用，而此项分布的方差等于其平均数。因此，当泊松分布的条件平均数作为一自变量方程而变化，其方差亦是如此。广义线性模型在 logistic 和泊松回归模型这种情形下，在模型的公式化及估计回归参数里，已明确地通过其概率分布体现了平均数和方差的关系。

经典回归亦假设在回归参数中模型是线性的，即其假设期望值或条件平均

数是回归参数的线性函数。例如，$E(y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$，或者 $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$。需注意在第二个模型中，参数是线性的但自变量是非线性的。事实上，经典线性回归是一个广义线性模型的特例，其因变量的条件平均数直接被模型化而没有对条件平均数进行某种转换。对其他的广义线性模型而言，条件平均数无法被写成回归参数的一个线性函数，但某些条件平均数的非线性方程则可以用参数的线性函数来表示，因此，称为广义线性模型。

一个简单的广义线性模型的例子是泊松回归模型(图 1.2)。所有广义线性模型的特性在这个例子中清楚可见。此外，也容易看出这个广义线性模型与经典回归模型之间的差异。

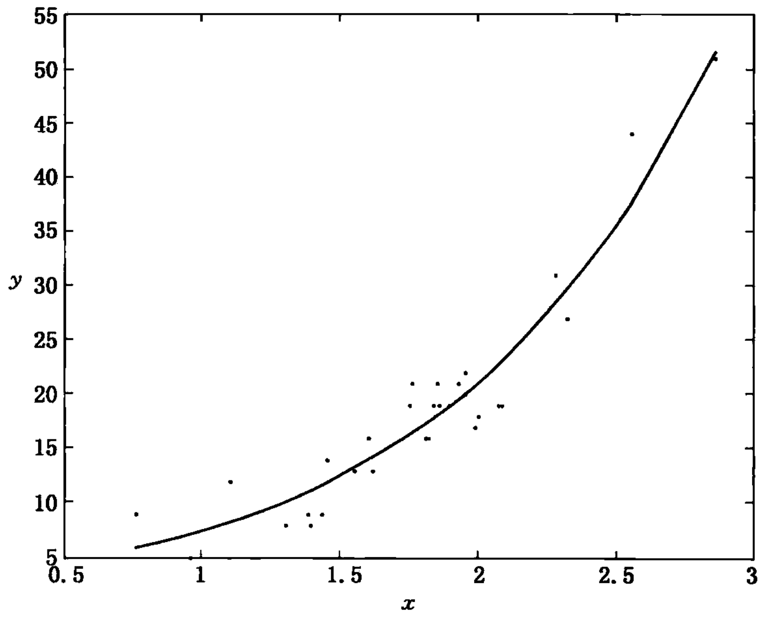


图 1.2 泊松回归模型

在这个泊松回归例子中，因变量的期望值或条件平均数为：

$$\lambda_i = e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}}$$

其中，λ_i 为对于每一个个体 i，泊松分布的条件式均数。它之所以为条件式的，

是因为该平均数根据回归参数而变动，包含固定的 β_j 以及随着分析单位（例如：个人）而改变的 X_j 的取值。我们要计算对于个体 i 的条件平均数 λ_i，需要代入他或她的自变量值，即 X_{ij}，X_{ij} 表示个体 i 第 j 个自变量的值，并估计未知常数的回归参数 β_j。具体的估计方法将留待后面讨论。我们必须使用最大似然法而非最小二乘法。

当因变量的分布是非正态的且它的方差为其平均数的一个方程，最小二乘估计值就不再如同它们在正态分布中一样，等于最大似然估计值了。在此类例子中，似然函数必须用适当的概率密度表示，以获取适合的参数估计值及其标准误。使用最小二乘会导致错误的参数估计值及标准误。

重点是条件平均数并非 β_j 的线性函数。如果我们对上述泊松回归模型的两边都取自然对数，就得到 $\log_e(\lambda_i)=\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}$。我们通过对条件平均数 λ（如 $\log_e(\lambda)$）执行一个非线性转换，将泊松分布的因变量与自变量之间的关系线性化。可见，$\log_e(\lambda)$ 被称为泊松回归模型的标准连结方程。它转换了因变量的条件平均数 λ，以至于转换后的值 $\log_e(\lambda)$ 为回归参数的一个线性模型。它之所以标准，是因为当 $\log_e(\lambda)$ 以指数型态表示时，它为泊松分布的自然参数。我们在后面也会看到，泊松变量的方差等于其平均数。因此当泊松分布的条件平均数增加时，与条件平均数相关的条件式方差也会随之增加。

目前已有些关于广义线性模型的好书（Fahrmeir & Tuz，1994；Le，1998；McCullagh & Nelder，1989；McCulloch & Searle，2001），但它们通常假设读者已有相当高程度的统计理解（sophistication）。本书仅假设读者具有基础的统计理解（inference）并对多变量回归比较熟悉。有关基础微积分及矩阵的知识并没有被假设，虽然此书的部分章节中会涉及。如果具备太少或不具备这方面知识的读者可以跳过或略读这些章节，也不会有太多的不连续性。此书的写作是非正式的，并且是从直观上来讨论重要的统计概念。目的是要告知读者相关的不同数据，并使其能选择适当的统计模型来分析数据及诠释结果。在附录中，我们还提供了如何运用 SAS 统计软件（SAS Institute，2002）的内容，以拟合本书所讨论的广义模型。

第 2 章 | 一些基础的模型化概念

我们在一般的多元回归分析的脉络下讨论统计模型的基本概念。假设因变量是连续分布的且对每一个观察值都有固定的方差，并且假设 y 的预测值，即条件平均数是回归参数的一个线性函数，那么，如果假设误差是正态分布的话，正常多变量回归模型就是特定广义线性模型中的一种。

如果有三个自变量，模型可以写成 $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$，其中 i 代表观察值，在绝大部分的应用中指单个人。假设 ε_i 有平均数 0 及固定的方差 σ^2。此外，假设 ε_i 与自变量不相关。此模型中系统性的(systematical)成分为 $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$，它是 y_i 的期望值，或者说，是对第 i 个观察值在给定 X_{i1}，X_{i2}，X_{i3} 的取值下其因变量的条件平均数。我们以此表示：

$$E(y_i \mid X_{i1}, X_{i2}, X_{i3}) = \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

此模型的随机成分为 ε_i。我们可以看到，当自变量改变时，条件平均数 μ_i 也会随之改变。相关的回归参数 β_1，β_2，β_3 表示每一个自变量与因变量 y_i 之间的相关性强度。一般而言，回归参数越大，相关变量与因变量之间的关系越强，假设自变量的方差大致相同。对于不同的观察值而言，β 是固定的，各个观察值间变化的是 X，并且影响条件平均数 μ_i 的改变。参数 β_0 被称为截距，是四个变量下多次原空间的 y 截距，即当所有自变量都为 0 时因变量的期望值。

自变量可以是连续性的、类别性的或者是两者的混合。它们也可以是 X 的转换，如 X^2 和 $\log X$，或如 X_1X_2 之类的交互作用，只要模型可以用线性的参数

型态所表示即可，例如：

$$E(y \mid X_1, X_2) = \mu_{y|X_1, X_2, X_1^2, X_2^2, X_1X_2}$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

作为类别变量的自变量

类别自变量可以通过指标(indicator)变量表示，这一点稍后会解释。指标变量定义了因变量的层级，或类别自变量各层级之间的差异，这些差异被显示出来以避免 $X'X$ 矩阵中重复的信息。我们由两个层级的类别变量的例子开始，然后再推广到有任意层级的类别变量。一个有两层类别自变量的例子是药物滥用的治疗变量，定义为两层——有治疗和没有治疗。我们可以通过 0—1 的指标变量量化这个变量，其中，1 代表这个人经历过治疗，而 0 则表示这个人没有经历过治疗。如果这是唯一的自变量且 y 是正态分布的，例如，药物滥用人群对于药物滥用治疗功效的信念，则 y 的条件平均数可以被表示成：

$$E(y \mid X_1) = \beta_0 + \beta_1 X_1$$

其中，如果该人经历过治疗，$X_1 = 1$，0 则表示其他。对于经历过治疗的人而言，$E(y \mid X_1 = 1) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$，而对于没有经历过治疗的人则是 $E(y \mid X_1 = 0) = \beta_0 + \beta_1(0) = \beta_0$。因此，$\beta_1$ 代表了有治疗和没治疗者之间的条件平均数的差异。β_1 有时会称为 X_1 对 y 的效应，在这个例子中，就是指治疗。而被编码为 0 的群体则被称为参照组。可以任意将某个群体标示为 1。注意只需要一个指标变量来表示该人归属于两类别中的哪一类。

使用指标变量可以推及三层甚至更多层级或类别的类别变量(例如，种族群体被分为白种人、非裔美国人、亚洲人及其他——这个类别变量有四个类别)。在这里有四个可能的指标变量：$X_1 = 1$ 表示白种人，0 则否；$X_2 = 1$ 表示非裔美国人，0 则否；$X_3 = 1$ 表示亚洲人，0 则否；$X_4 = 1$ 表示其他种族类别，0 则否。要决定一个人的种族类别只需要这四个指标变量的其中三个(如果我们知道一个人四个指标变量的其中三个的值，则另一个指标的值就已经被自动赋

予了)。如果我们将所有四个指标都放入回归模型中,则作为一个组(set)而言,它们是多余的,且从个人样本中无法估计与其相关的回归参数。故可以任意丢弃其中一个种族指标变量,而被丢弃的那一个指标变量就成了解释剩余三个指标变量的回归参数的参照组。

如果我们想要检验种族对于呈正态分布的成就测验成绩的效应,就要将种族定义成前述的类别变量。我们舍弃 X_4 这一代表其他种族群体的指标变量,则可以将控制了种族类别的 y 的条件平均数表示为:

$$E(y \mid X_1, X_2, X_3) = \mu_{y|X_1, X_2, X_3} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

对白种人而言,$X_1 = 1$, $X_2 = 0$, $X_3 = 0$,所以其期望值为 $\mu_{y|X_1=1, X_2=0, X_3=0} = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1$。

同样的,非裔美国人的 Y 的条件平均数为 $\mu_{y|X_1=0, X_2=1, X_3=0} = \beta_0 + \beta_2$,亚洲人的则为 $\mu_{y|X_1=0, X_2=0, X_3=1} = \beta_0 + \beta_3$。因为其他种族类别在这三个指标变量都取 0 值,其条件平均数为 $\mu_{y|X_1=0, X_2=0, X_3=0} = \beta_0$。因此,白种人的条件平均数 $\beta_0 + \beta_1$ 与其他种族群体条件平均数 β_0 的差异为 β_1,即 β_1 为白种人与其他种族,也就是与参照组的平均数的差异。β_2 和 β_3 分别为非裔美国人和亚洲人与其他种族的平均数的差异。我们可以通过 β_1, β_2 和 β_3 之间的不同组合的相减,来获得不同种族群体之间的差异。例如,非裔美国人平均数与白种人平均数的差异为($\beta_1 - \beta_2$)。注意,在这个模型中,有四个种族群体及四个回归参数,β_0、β_1、β_2 和 β_3。含有类别自变量的模型,其回归参数的个数不可能超过用来定义类别变量的自变量数目。

我们再看另一个稍微复杂点的两个类别自变量的回归模型,每个类别自变量都有两个层级。假设因变量 y 是一个测量药物滥用者治疗满意度的呈正态分布的变量。自变量为(两个)指标变量,X_1 取 1 表示病患属于住院计划(residential program),取 0 则表示门诊计划(outpatient program),而 X_2 取 1 表示病患是男性,取 0 则为女性。主效应模型可以用 $\mu_{y|X_1, X_2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ 表示,其中 β_1 为住院计划(residential program)患者和门诊计划(outpatient program)

患者的条件平均数的差异。同样，β_2 为性别效应，表示男性和女性之间的差异。而这两个差异，即参数 β_1 和 β_2，对其他遗留变量的效应都进行了调整。也就是说，β_1 对于 X_2 的效应调整过了，而 β_2 对于 X_1 的效应也调整过了。

我们可以通过加入一个反映性别和这两种治疗变量交互作用项的参数改进这个模型。这种类型的交互作用可以通过两个指标变量的乘积项而形成，同时产生第三个变量，X_1X_2，它也是一个 0—1 变量。当它等于 1 时，代表患者是一名男性且属于住院治疗计划(residential treatment program)，0 则否。此模型就变成 $\mu_{y|X_1,X_2,X_1X_2}=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$。现在 X_1 对 y 的效应取决于 X_2 的层级。如果 $X_2=0$，则 $\beta_3X_1X_2=0$，且 X_1 的效应为 β_1，因为与 X_1 有关的唯一项为 β_1X_1。如果 $X_2=1$，则 X_1 对 Y 的效应为 $(\beta_1+\beta_3)$，因为在这个模型中，$\beta_1X_1+\beta_3X_1=(\beta_1+\beta_3)X_1$。

回归模型的必要成分

这里，我们总结了回归模型的必要成分。首先，我们选择一个因变量 y，假设其为一组自变量的一个方程。自变量的选择是根据研究目的及研究者对于研究领域的实际知识来决定的。一个统计学家在这方面可能没有太大帮助，然而，他可以就研究设计提出建议，特别是样本数、用来选择研究参与者的概率模型，以及广义线性回归模型的种类。

其次，离差是对回归模型拟合优度的一种测量。每一个广义线性模型都有关于它本身的一个特定离差。对于以正态分布为基础的回归模型而言，就是大家所熟悉的误差平方和。关于离差将会在后面做更加详细的讨论。

通常，我们最初的一组自变量包含了那些回归参数不具有统计显著性的自变量的集合。通过比较各模型的拟合，我们可以减少此模型中的参数数目。这个模型一旦被确定，我们的兴趣则集中于回归参数的估计值及其估计标准误。回归参数估计值表示各自变量在解释因变量的变异上的相对重要性。

总之，对广义线性模型而言，我们需要指明因变量的概率分布。到目前为

止，我们已经讨论过正态分布和泊松分布，但还有其他的分布形式，例如二项分布，我们将留待后面做详细讨论。另一个需要我们指明的重要成分是回归方程，它设定了条件平均数如何与自变量相关。我们讨论过在正态和泊松分布下因变量的回归方程，我们将在后续内容讨论其他形式。以概率分布形式呈现误差分布的形式，如正态、二项、泊松及其他，以及回归方程的形式，对于指出适当的对数似然函数是必要的。对数似然函数等同于概率密度函数的对数，然而在前者中，样本数据被视为固定的且参数为变量，但在后者中，参数则被视为固定的且数据会变化。样本的对数似然函数会被用来获取最大似然回归参数估计值及它们的标准误。除了正态回归模型的最大似然估计值外，最大似然估计值因为方程系统的复杂性，不能以标准分析方法解出，而需用迭代加权最小二乘法(iterative reweighted least squares)的计算机演算来执行参数估计。

第 3 章 | 经典多元回归模型

多元回归分析为一种广义线性模型,其条件平均数是回归参数的一个线性函数。这跟那些条件平均数的函数是回归参数的一个线性函数的广义线性模型有所不同。平均数的函数 $g(\mu)$被称为连结函数。例如,泊松分布的连结函数为 $\log_e(\mu)$,可表示为回归参数的一个线性函数,亦即 $\log_e(\mu) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j$。对正态分布而言,连结函数即为同一性(identity)函数 $g(\mu) = \mu$。多元回归可被写成:

$$E(y \mid X_1, X_2, \cdots, X_p) = \mu = \beta_0 + \sum_{j=1}^{p}\beta_j X_j$$

如果我们将一个有关截距参数 β_0 为 1 的列向量(column vector)包含进来,则模型可以更完整地表示为:

$$E(y \mid X_0, X_1, X_2, \cdots, X_p) = \mu = \sum_{j=0}^{p}\beta_j X_j$$

其中 $X_0 = 1$。误差 $y - \mu = \varepsilon$ 被假设是正态分布且有固定方差的。

我们通过查特吉和普赖斯(Chatterjee & Price, 1977)的一个多元回归分析实例来讨论广义线性模型。此根据来自一个大金融组织的职员所做的调查数据。该调查问卷包含职员对其主管的满意度。其中一个问题是有关主管整体表现的测量。另外的问题则涵盖了职员与其主管在特定活动中的互动。此研究目的是要解释特定的主管特质与职员对主管的整体满意度之间的关系。

在这个例子中,有六个问题被选择当做可能的解释变量。因变量是职员对主管的整体工作表现的评分。它是一个 5 分的尺度,从 1(非常满意)到 5(非常不满意)。而六个自变量也是以 5 分的尺度评价主管的行为。X_1 处理职员抱

怨;X_2 不允许特权;X_3 学习新事物的机会;X_4 根据绩效加薪;X_5 对不良表现太过挑剔;X_6 升迁到更好职位的速率。这些自变量可以做如下分类:一组是有关职员和主管间直接人际关系的变量 X_1,X_2 和 X_5,另一组则是考虑工作整体而非人际关系的 X_3 和 X_4 两个变量,剩下的变量 X_6 不牵涉对主管的评价,而是职员对其自身在公司中升迁看法的一般性测量。数据从该组织所有部门中随机挑选 30 个部门收集而来,每一个部门大概都有 35 名职员及一名主管。

假设与模型方法

注意,因变量只会取五个值且在这五个值上的分布是偏斜的,因为大部分主管较容易获得好评而非劣评。因此这违反了多元回归分析中的普遍假设:因变量是一连续变量且为正态分布的。即使背离从正态性出发的假设,只要因变量有数目相对大的值且其分布相对来说对称,则此模型仍然是稳健的。

基本多元回归模型亦假设横跨各观察值的因变量,在此例子中指的是职员,是各自独立的。这个假设也可能被推翻,因为同一部门中的职员彼此之间可能相同,即同一部门内各职员对于因变量的回答可能会彼此相关,因为他们都受到其部门的同样一组影响。然而,我们不会期望不同部门职员的因变量回答是相关联的。在同一群体或聚类(cluster)中的个人回答之间的相关被称为组内(intraclass)相关。基本多元回归模型可以被修正,以说明在回归参数及其标准误的估计中所存在的组内相关。回归模型通过增加随机成分来修正,并且被视为混合效应模型。这类模型在此书中不再进一步讨论。

需要特别注意的是,对于自变量的分布并没有任何假设。它们可以是连续的、离散的、高度偏斜的或彼此相关的。因此,在主管表现的例子中,自变量并没有问题,即使它们是离散的甚或是高度偏斜的。对于任何一种回归模型来说都是如此。

上述例子中所存在的问题是:关于因变量的非正态分布及在相同部门内部各职员回答的组内相关。然而,查特吉及普赖斯(Chatterjee & Price, 1977)通

过汇总 30 个部门间的个人层级数据及在多元回归分析模型中使用部门作为分析单位,改善了他们的研究实例中所存在的这两个问题。因为他们对部门主管的整体绩效预测感兴趣,所以是合理的。对于 30 个部门来说,每个部门都只有一名主管,故通过分别在每一个主管所在的部门中,汇总该部门内约 35 名职员对于六个变量的看法,进而预测该主管的表现是合理的。相对于简单地在各个部门内选取一些职员来对其主管的整体表现评分,这个方法更加可信。这是因为职员的偏误及评分误差在大样本的评分者中易于被抵消。然而,要注意到汇总各个部门间的数据有可能会遗失部门间每个人回答变异的信息。

作者利用下述步骤汇总个人层级的数据至部门。为了反映六个自变量和一个因变量,针对这七个项目分别产生一个二元变量,即把(原来的)5 分回答尺度重新组合为两类:好评及差评。好评为 1(非常满意)或 2(满意),差评则为其他剩余的回答 3、4 或 5。对 30 个部门都计算出七个项目好评的比例。因此,我们可将得到的代表 30 个部门对其主管整体表现好评比例的列向量作为因变量,并且获得一个代表 30 个部门六个自变量的好评比例的 30(列)×6(行)的矩阵。以这 30 个部门为分析单位的数据被用来估计六个回归参数及其标准误。

我们将在后面看到,由于因变量是对主管整体表现好评的概率,logistic 回归分析可以被用来模型化六个自变量与主管表现之间的关系。logistic 回归被用来模型化一个特殊反应发生的概率——在这个例子中,就是对主管的好评。logistic 函数为非线性的。然而,大多数部门的主管好评比例(概率)都落在 0.40 到 0.60 的区间中,在此区间 logistic 函数大致是线性的。因此多元回归模型在对这种数据进行建模时可能是最佳的选择,且较 logistic 回归模型更容易解释。

回归分析结果

多元回归分析的结果见表 3.1,该表指出只有 X_1(关于主管如何处理抱怨的满意度)是高度显著的,并且它有最大的回归参数估计值。回归参数 0.613 表

示，在满意其主管处理抱怨的职员百分比中每增加一个单位，就会使主管整体好评增加 0.613 个百分点。也就是说，X_1 增加 1%会造成 y 增加 0.613%；X_1 增加 10 个百分点则会造成 y 增加 6.13 个百分点。这表明，当其他五个自变量保持不变时，X_1 和因变量之间有一种强关系。数值次大的回归参数 X_3（学习新事物机会的满意度）为 0.320。虽然它在 0.10 水平上统计显著，但在传统的 0.05水平上并不显著。

表 3.1　预测正面主管评价的回归参数估计值

变　量	回归参数	标准误	t 比率	显著程度
X_1（抱怨）	0.613	0.1610	3.81	0.001
X_2（特权）	−0.073	0.1357	−0.54	NS
X_3（学习）	0.320	0.1685	1.90	0.07
X_4（加薪）	0.081	0.2215	0.37	NS
X_5（挑剔）	0.038	0.1470	0.26	NS
X_6（工作升迁）	−0.217	0.1782	−1.22	NS
截距	10.787	11.5890	0.93	NS

注：$R^2 = 0.7326$；残差标准差 = 7.068；NS = 在 0.05 水平上不显著。

多元相关

多元回归相关（R^2）0.7326 代表自变量解释因变量变异的比例。它被定义为：

$$R^2 = 1 - \frac{error\ sum\ of\ squares}{total\ sum\ of\ squares}$$

其中，误差平方和（*Error Sum of Squares*，*ESS*）被定义为 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$，$\hat{y}_i$ 为根据回归模型对 y_i 的预测值，即 $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_{ij}$。总平方和（*Total Sum of Squares*，*TSS*）被定义为 $\sum_{i=1}^{n}(y_i - \bar{y})^2$，$\bar{y}$ 是因变量的平均数。如果 *ESS* 比 *TSS* 小，则 R^2 会更高。多元相关 $R = \sqrt{0.7326} = 0.856$ 可以

被定义为 y 和 $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$ 之间的相关。

假设检验

通常，最初被应用的一个标准检验是用于检验虚无假设，即所有回归系数都为零。在我们的例子中，虚无假设为 $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$。注意，我们不将截距参数 β_0 包含在检定中，是因为我们只关注六个自变量与因变量之间的关系。

要检验虚无假设，即一组回归参数等于零，我们可以通过估计两个回归模型的误差平方和开始。一个包含全部回归参数的模型被称为完整模型（*Full Model*，*FM*），在我们的例子中，即要包括所有六个自变量。另一个模型被称为简化模型（*Reduced Model*，*RM*），它去除了参数被假设为零的自变量。整体检验的简化模型，即六个自变量和主管评分间没有关系，要去除六个自变量并且只在简化模型中保留截距。简化模型中的截距即为因变量的平均数 $\bar{y}$。接着，我们要计算两种模型的误差平方和 $ESS(FM)$ 和 $ESS(RM)$，计算差异 $ESS(RM) - ESS(FM)$，并且用它除以完整模型和简化模型之间的参数数目——分别以 p_f 和 p_r 表示差异。在我们的例子中，$p_f = 7$，$p_r = 1$。这个比率如下所示：

$$\frac{ESS(RM) - ESS(FM)}{p_f - p_r}$$

即为一个 F 比率的分子。$ESS(RM)$ 至少会和 $ESS(FM)$ 一样大，因为它用较少的自变量去预测 Y——在我们的例子中，简化模型中不包含任何自变量。如果这个取值为正数的差异很小，则表示简化模型同完整模型一样适合该数据。我们需要一个项目测量这个比率的差异。这个项目形成 F 比率的分母，即为单纯将 $ESS(FM)$ 除以其自由度，也就是样本数目 n 减掉在完整模型中估计的回归参数数目。在我们的例子里，$(n - p_f) = 30 - 7 = 23$。

因此，F 比率为：

$$\frac{\frac{ESS(RM)-ESS(FM)}{p_f-p_r}}{\frac{ESS(FM)}{n-p_f}}$$

分子项的自由度为 $(p_f - p_r)$，而分母项的自由度为$(n - p_f)$。在虚无假设下简化模型外的参数为零，F 值为具有前述分子和分母自由度的分布。我们通过数据中计算的 F 值以及分子和分母自由度，可在 F 表中查询相应的值。查特吉和普赖斯(1977)计算关于虚无假设，即所有回归参数都为零，或者同意义的，在六个自变量与整体主管绩效评分间没有线性关系的 F。

对于这个 F 检验，其成分为 $ESS(FM) = \sum_{i=1}^{30}(y_i - \hat{y}_i)^2 = 1149$；$ESS(RM) = \sum_{i=1}^{30}(y_i - \hat{y}_i)^2 = 4297$；$p_f - p_r = 7 - 1 = 6$；以及 $n - p_f = 30 - 7 = 23$。这会产生一个 F 比率：

$$\frac{\frac{4297-1149}{6}}{\frac{1149}{23}} = 10.5$$

分子和分母的自由度分别为 6 和 23。在一个有着 6 和 23 的自由度的 F 表中查询这个 F 值，显示它在 0.001 的水平上显著。因此，我们可以拒绝所有回归参数都等于零的虚无假设，并得出结论，相对于简化模型 $E(y) = \mu = \beta_0$，完整模型 $E(y \mid X_1, X_2, \cdots, X_p) = \mu = \beta_0 + \sum_{j=1}^{p}\beta_j X_j$ 更好地拟合了数据。

得出上述结论是因为两个回归参数和 X_1 在 0.001 水平上都是显著的。笔者通过画出标准化残差(y 轴) 及拟合或预测值(x 轴)来检查模型假设的违反或模型的错误设定。标准化残差为 $y_i - \hat{y}_i$ 除以误差或残差的标准差，即：

$$\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-p_f}}$$

如果模型的设定正确，则标准化残差的散点图应该显得较为随机，没有系统性的型态，且 95%的残差应落于−2 和 2 之间，或者是在残差平均数(假设下为零)的两个标准误范围内。它们的残差图显示没有模型设定错误的证据。

查特吉和普赖斯(1977)也勾画了标准化残差和最重要的自变量 X_1 的图形。散点图看起来是随机的,没有大的标准化残差,并且没有证据显示任何系统性的弯曲。弯曲可表示在回归模型中需要加入 X^2 这一项目。

回到我们的完整模型,相当清楚的是,只有 X_1 和 X_3 看起来是重要的。因此,检定 $\beta_2=\beta_4=\beta_5=\beta_6=0$ 的虚无假设看似很合理,以决定我们是否可以简化这个完整模型。

为了检验这个假设,我们用前述相同的 F 检验过程,比较完整模型 $\mu=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6$ 和简化模型 $\mu=\beta_0+\beta_1X_1+\beta_3X_3$。

因此,我们再次计算 $ESS(FM)$, $ESS(RM)$, p_f-p_r 和 $n-p_f$,并且将它们代入 F 比率的公式中。可以得到:

$$F=\frac{(1254.6-1149)/4}{1149/23}=0.528$$

有着 4 和 23 的自由度并且有一个很小的、在 0.05 水平上不显著的 F 值。因此,我们接受虚无假设,即四个变量 X_2、X_4、X_5 和 X_6 并不需要包含在这个模型中,从而接受较简单的简化模型 $E(y\mid X_1, X_3)=\beta_0+\beta_1X_1+\beta_3X_3$,因为它优于较复杂的完整模型 $E(y\mid X_1, X_2, X_3, X_4, X_5, X_6)=\beta_0+\sum_{j=1}^{6}\beta_jX_j$。

此简化模型的 R^2 为 0.7080,较完整模型的($R^2=0.7326$)稍小。X_1 的回归参数最为重要,在简化模型中为 0.643,而在完整模型中为 0.613。注意,这个模型中的所有自变量都是连续性变量。反映部门归属的类别自变量也可以被包含在模型中。例如,我们可以有一个 0—1 指标变量,1 代表部门执行会计功能,而 0 表示其他。我们也可以通过纳入假设互相作用的变量间适当的乘积项,在模型中加入交互作用项。然而,我们必须谨记,在只有 30 个观察值的情况下,我们不能拟合很多参数。否则,我们将冒过度拟合的风险,也就是参数过多并且超出了数据所能支持并导致合理推论的范围。在极端的例子中,如果我们的模型包含了 30 个参数,则模型完全可以拟合数据,然而,我们并不能凭借和观察值一样多的参数来简化任何事。这会造成 ESS 没有自由度,且导致无法进行备择假设检验。即使 $ESS(FM)$ 有着很小的自由度,F 检验仍将有低效力。要建

立一个好的回归模型，需要从指明一个因变量及一组自变量开始，并被一组良好阐述的假设所驱动，而这又依赖对研究主题的认知。研究者随后才能估计模型参数并衡量拟合优度。最初的完整模型则会通过假设检验而去除一些参数予以修正。回归诊断，如残差图，有助于模型的其他修正，比如增加 X^2 项或交互作用项。如果是时间序列数据，则残差可能会彼此相关，因此修正的模型必须要能考虑到在回归参数估计中的相关残差且对它们做各式各样的假设检验。

第 4 章 | 广义线性模型的基本原则

第 3 章已讨论了我们都熟悉的一个广义线性模型，也就是经典多元回归模型。这个模型可以被推广至其他情况，即因变量是离散的、非正态分布的，且其方差取决于其平均数。

广义线性模型牵涉到以一组自变量或称协变量的线性函数来预测因变量的条件平均数或条件平均数的某种函数。也就是说，对于每个观察值或研究对象，其期望值或因变量的期望值的某种函数，是根据其自变量或协变量而定的。除了正态分布外，广义线性模型的误差方差是其平均数的一个函数。例如，一个 0—1 二元变量有平均数 π，表示事件 1 发生次数的比例，且其方差为 $\pi(1-\pi)$。要估计回归系数及其标准误，我们需要指明误差项的概率分布，由此我们可以指明适当的似然函数并用该似然函数解出回归参数。

广义线性模型可以处理因变量的条件平均数为回归参数的非线性函数和因变量为非正态分布的数据。广义线性模型的两个成分为连结函数与误差分布。连结函数是因变量平均数的转化，而此转化的变量为回归参数的一个线性函数。例如，泊松回归模型的连结函数为 $g(\mu)=\log_e(\mu)$，所以因变量$g(\mu)$是与自变量相关的回归参数的一个线性函数，也就是 $\log_e(\mu)=\sum_{j=0}^{p}\beta_j X_j$。注意，$g(\mu)$是一个回归参数的非线性函数，因为对方程两边都取幂会导致：

$$\mu=e^{\sum_{j=0}^{p}\beta_j X_j}$$

此泊松分布的对数连结函数也被称为标准连结，因为它是当泊松分布以指数形式表示时，变成标准参数 θ 的 μ 的转化；也就是 $g(\mu)=\theta=\log_e(\mu)$。此连结函数最常被用在泊松回归中，虽然其他的连结函数也是可能的。例如，我们可以

用非标准的同一性(identity)函数 $g(\mu)=\mu$。在有些例子中，一个非标准的连结可能比标准连结更好地拟合了某种特别的数据。我们在后面会看到，广义线性模型假设因变量分布是指数家族中的一员。当分布以指数形式表示时，每个指数家族中的分布都有其自己的标准参数 θ 为其平均数的函数。当然，函数$\theta(\mu)$对不同指数家族中的成员都不同。如泊松分布，我们知道它是 $\theta(\mu)=\log_e(\mu)$。有关 logistic 回归的二元分布则是：

$$\theta(\mu)=\log_e\frac{\mu}{1-\mu}$$

对于正态分布，$\theta(\mu)=\mu$——也就是说，它是一个同一性连结。

除了正态分布以外，广义线性模型的第二个成分，是因变量的方差为其平均值的一个函数。这是指数家族成员的一个分布特性，也是广义线性模型背后的响应(response)分布。泊松分布的方差为 $\mathrm{Var}(y)=\mu$，而二元分布的方差为 $\mathrm{Var}(y)=\mu(1-\mu)$。对正态分布而言，方差是固定的，即 $\mathrm{Var}(y)=\sigma^2$。

广义线性模型假设因变量的观察值 $y_1, y_2, \cdots, y_n$ 是相互独立的，且共享指数家族中相同形式的参数分布。观察值的平均数 $\mu_1, \mu_2, \cdots, \mu_n$ 可以不同，但每个观察值必须都能由相同的概率分布所产生(例如，都以泊松分布产生)。这意味着每个观察值的平均数不同，因为广义线性模型假设平均数或平均数的某种非线性函数与一组自变量有关。也就是说，假设我们有一组$(p+1)$回归参数 $\beta_0, \beta_1, \cdots, \beta_p$，且有一组相关的自变量 $X_1, X_2, \cdots, X_p$，则适当的连结函数为 $g(\mu)=\sum_{j=0}^{p}\beta_j X_j$。

其次，我们检视广义线性模型所立足的、来自于指数家族概率分布的特性。

指数家族分布

广义线性模型涉及可以用指数形式来表示的概率分布，这些分布为指数家族分布的成员。当以指数形式表示时，有一个标准参数为平均数的函数，且方差也是平均数的一个函数。例如，泊松分布的标准参数为 $\log_e(\mu)$，分布

的方差为 μ。

正态分布通常表示为：

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

其中，μ 和 σ^2 分别为分布参数的平均数和方差。注意，分布已是部分为指数形式。使用代数方法，可以表示为：

$$f(y \mid \mu, \sigma^2) = e^{\frac{\left(y\mu - \frac{\mu^2}{2}\right)}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log_e(2\pi\sigma^2)\right)}$$

所有指数家族的分布都可以表示为：

$$f(y \mid \theta, \varphi) = e^{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)}$$

θ 被称为标准或自然参数，为分布平均数(μ)的一个函数；$b(\theta)$ 为标准参数的一个函数，也是平均数的一个函数，因为 θ 为平均数的函数；ϕ 为离散参数，扮演着定义 y 的方差的角色；而 $c(y, \phi)$ 为观察值及离散参数的函数。通过等化以指数形式表示 μ 和 σ^2 的正态分布 $f(y|\mu, \sigma^2)$ 的项目，与以标准参数 θ 和 ϕ 表示的指数型态中的项目，我们便可以确定对正态分布而言，这些参数(θ 和 ϕ)及函数 $b(\theta)$ 和 $c(y, \phi)$ 的含义。我们发现 $\mu = \theta$, $b(\theta) = \theta^2/2$, $\phi = \sigma^2$，且：

$$c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log_e(2\pi\sigma^2)\right)$$

一个重要的成分 θ 为 μ 的函数，以 $\theta(\mu)$ 表示，被称为标准连结函数。它连结了平均数和标准参数，并可以用回归参数的一个线性函数来表示。另一个指数家族分布的重要成分是方差函数，也就是 $b(\theta)$ 的二次导数 $b''(\theta)$。对正态分布而言，二次导数为 $b(\theta) = \theta^2/2 = 1$。此分布的方差为 $\varphi b''(\theta)$，φ 为离散参数，且 $b''(\theta)$ 为 $b(\theta)$ 的二次导数。对于正态分布，$\varphi = \sigma^2$ 且方差函数 $b''(\theta) = 1$，故正态分布变量的方差为 σ^2。它是一个常数且不是平均数的一个函数。

每个指数分布家族中的成员都有其自身的连结函数 $\theta(\mu)$ 和方差函数 $b''(\theta)$。方差函数也可以用平均数 μ 来表示，因为 θ 为 μ 的函数。以 $V(\mu)$ 来表示，它说

明方差为 μ 的一个函数。我们如果检验泊松分布并以指数形式来表示，故可以确定 $\theta(\mu)$和 $b''(\theta)$。泊松分布有一个参数 μ(也常以 λ 表示)。泊松分布以先前讨论的平均数表示为：

$$f(y \mid \mu) = \frac{\mu^y e^{-\mu}}{y!}$$

"!"记号为阶乘，且 $y! = y(y-1)(y-2)(y-3)\cdots 1$。泊松分布的指数形式可被表示为 $f(y \mid \mu) = e^{y\log_e\mu - \mu - \log y!}$。

将此与标准指数形式 $f(y|\theta, \phi)$等化，我们得到 $f(y|\theta, \phi) = e^{y\theta - e^{\theta} - \log_e y!}$。因为 $\theta = \log_e \mu$，$b(\theta) = e^{\theta}$，$\phi = 1$ 且 $c(y, \theta) = \log_e y!$。此标准连结函数为 $\theta = \log_e(\mu)$，方差函数为 $b''(\theta)$。用一个 μ 的函数表示的方差函数是同一性(identity)函数 $V(\mu) = \mu$，因为 $e^{\theta} = e^{\log_e \mu} = \mu$。因为 $\phi = 1$，平均数为 μ 的泊松分布，其方差也为 μ。泊松分布不像正态分布那样涉及一个未知的离散参数，因为它是一个等于 1 的常数。这里说明的原则可以应用到指数家族的其他分布。对所有的例子而言，标准参数可以被定义为自变量的唯一线性函数，即 $\theta = \sum_{j=0}^{p} \beta_j X_j$。

经典正态回归

对正态分布的因变量而言，我们通常使用同一性连结——也就是 $g(\mu) = \mu$。注意，正态分布的标准参数为 $\theta(\mu) = \mu$，所以 $g(\mu) = \mu$。我们直接模型化 μ 而不用对 μ 做任何转换。因此 $y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$，$y$ 的期望值或 μ 为 $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$，且假设 ε 服从均值为 0、方差为 σ^2 的正态分布。ε 的方差并不像其他的广义线性模型那样根据平均数而定；对于所有的观察值而言，它(方差)被假设为是固定不变的。此模型的回归参数可以被解释为，在其他自变量固定的条件下，相对应的自变量中每一单位的增加所导致 μ 增加的效应。

对于其他广义线性模型，如 logistic 和泊松回归模型，回归参数表示经由连结函数(也就是，平均数的某种非线性转换，且并不是平均数)相对应自变量的每一单位的增加所造成的增加效应。

logistic 回归

logistic 回归被用来模型化某事件的概率，比如，以一个学生特征函数表示的退学。logistic 回归的标准连结为 logit，即$\log\frac{\pi}{1-\pi}$，其中 π 为二元因变量的平均数或事件发生的概率。因此，$\log\frac{\pi}{1-\pi}$ 可表示为自变量的一个线性函数，即$\log\frac{\pi}{1-\pi}=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$，其中 β_0，β_1，…，β_p 为回归系数。在似然函数中运用的误差分布为二项式分布。对二项式分布而言，方差为平均数的函数，它等于 $\pi(1-\pi)$。虽然这是对于一个 logistic 回归模型的典型设定，但也还有其他设定方式。例如，可以运用 probit 连结函数，但如果有证据显示较二项式分布能更多地解释二元因变量的变异的话，则可以运用 beta 二项式误差分布。

对于 logistic 回归模型，$\log\frac{\pi}{1-\pi}=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$，所以 β_j 对 logit 的效应是叠加的。由于难以解释对数发生比(log of odds ratio)的叠加效应的大小，因此，我们通常对前述的方程两边指数化可得到：

$$\frac{\pi}{1-\pi}=e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}=e^{\beta_0}e^{\beta_1X_1}\cdots e^{\beta_pX_p}$$

我们可从此转换看出，被指数化的回归参数 e^{β_j} 现在表示，在其余自变量固定不变的情况下，X_j 每一单位的增加所导致的发生比的倍数效应。而 e^{β_j} 项表示发生在 X_j+1 值上的二元结果的发生比除以发生在 X_j 值上的发生比。注意：我们是模型化 $\frac{\pi}{1-\pi}$ 而非 π。

比例风险生存模型

生存模型牵涉以自变量的函数来模型化一个事件发生的时间(例如，死亡、退学或找工作)或某事件发生的时间的某个函数。我们将模型化一个以时间为自变量来表示风险的风险函数 $h(t)$。风险是在一给定时间内某事件发生的即

时概率。一个常用的生存时间数据的模型为 Cox 比例风险模型，定义为 $h(t)=h_0(t)e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}$，其中 $h_0(t)$ 被称为在时间 t 的基线。它是在没有协变量下的风险函数。如果我们将两边除以 $h_0(t)$，就得到：

$$\frac{h(t)}{h_0(t)}=e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}$$

可显示一比例词从何而来。对于每一个人而言，横越不同时点的 $e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}$ 是固定的，也显示在每一个 t 值下，任何个人的风险函数是基线风险的一个固定比例。对 Cox 比例模型两边取对数，则风险的对数可以用自变量的一个线性函数来模型化。

第 5 章 | 最大似然估计

最大似然估计基于估计参数——在我们的例子里是估计回归参数——应该是那些可以极大化说明样本数据的密度函数值的参数。也就是说，根据样本数据，最大似然估计会找到最可能产生样本观察值的参数值。当在给定数据下以一个参数的函数表示此概率密度即称为似然函数。密度函数和似然函数是相同的，但前者视参数是固定的且数据是变化的，后者则视数据是固定的而参数是变化的。最大似然估计值为那些最大化似然函数的参数估计值。在一些例子中，它们可以通过偏微分分析出来。在较为复杂的例子中，可能无法直接分析求解，而必须使用计算机运算。最大似然估计值有很强的统计特性，如有效性。

最大似然估计需要指明假设被用来描绘样本数据的一个概率密度函数。由于正态回归模型概率函数的特殊数学形式，回归参数的最大似然估计值等同于最小二乘估计值。但是，对于其他广义线性模型来说并非如此。

我们已讨论过的回归方程需要凭借一个由随机样本估计得出的总体参数。为此，我们需要一个统计模型，指出数据是如何产生出来的。对于正态分布的数据，正态分布随机变量 y 的密度函数为：

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

第 i 个样本成员的 y 的观察值为 y_i，且假设其分布为：

$$f(y_i \mid \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}}$$

即对于每一个观察值 y_i，假设是从一个平均数为 μ_i 的正态分布中产生的，但每一个观察值有着固定的方差 σ^2，因为在 σ^2 上没有小标 i。这是符合正态回归假设的，即条件平均数在不同个人间有所不同，为一协变量函数，但 y_i 的方差则固定不变。

现在，假设 μ_i 为回归参数的一个线性函数，则该回归模型可以表示为：

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

因此，我们把 y_i 的密度表示为：

$$f(y_i \mid \beta_0, \cdots, \beta_p, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}}$$

为了节省空间，我们将 $\beta_0, \cdots, \beta_p$ 集体表示为一个列向量 β。因为 y_i 被假设为是相互独立的，样本观察值的联合分布可表示为：

$$f(y_1, y_2, \cdots, y_n \mid \beta, \sigma^2) = f(y_1 \mid \beta, \sigma^2) f(y_2 \mid \beta, \sigma^2) \cdots f(y_n \mid \beta, \sigma^2)$$

其中 n 为样本数。因此：

$$\begin{aligned} f(y_1, y_2, \cdots, y_n \mid \beta, \sigma^2) = & \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}} \\ & \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}} \cdots \\ & \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}} \end{aligned}$$

可以被缩写为：

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}}$$

其中 $\prod_{i=1}^{n}$ 表示 n 个概率密度的乘积。因为常数 $\frac{1}{\sqrt{2\pi\sigma^2}}$ 自乘 n 次，联合概率密

度的一个因素为：

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n = \frac{1}{(2\pi\sigma^2)^{n/2}}$$

其中指数 1/2 表示取平方根。另一个牵涉 n 个指数的乘积的因素为：

$$e^{-\frac{1}{2}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}}$$

由于指数可以被加总，

$$e^{-\frac{1}{2}\sum_{j=1}^{n}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}}$$

因此：

$$f(y_1, y_2, \cdots, y_n \mid \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{1}{2}\sum_{j=1}^{n}\frac{(y_i-(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}}$$

以矩阵形式可以写成：

$$f(y \mid \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}}$$

其中，X 是一个对 p 个自变量观察值的 $n(p+1)$ 矩阵，且有一前导的列向量(leading column vecter)1 对应于截距参数 β_0，即为列向量 β 的第一个要素。

联合概率函数 $L(\beta, \sigma^2 \mid y)$为给定参数$(\beta_0, \beta_1, \cdots, \beta_p)' = \beta$的条件下随机变量$(y_1, y_2, \cdots, y_n)' = y$ 的一个函数。为了推论 β，我们将从样本中产生的 y 值视为固定的，且 $f(\beta, \sigma^2 \mid y)$为 β 的一个函数。此以 $L(\beta, \sigma^2 \mid y)$而非 $f(\beta, \sigma^2 \mid y)$表示。我们称之为似然函数。注意，$f(y \mid \beta, \sigma^2) = L(\beta, \sigma^2 \mid y)$。

我们可以估计最大化似然函数 $L(\beta, \sigma^2 \mid y)$的回归参数 β，以 $\hat{\beta}$ 表示。这些估计值被称为最大似然估计值(*MLE*)。它们是最可能从样本观察值 $y = (y_1, y_2, \cdots, y_n)'$ 中所产生的回归参数的估计值。

因为似然函数中有指数项，故比较容易处理似然函数的对数。由于对数函数是一个单调函数，最大化 $L(\beta, \sigma^2 \mid y)$的 β 值就等同于最大化 $\log_e L(\beta, \sigma^2 \mid y)$

的 β 值。在 $L(\beta, \sigma^2 | y)$ 的两边取对数，我们得到：

$$\log_e L(\beta, \sigma^2 \mid y) = -\frac{n}{2}\log_e(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij}))^2}{\sigma^2}$$

或者，以矩阵形式来表达：

$$\log_e L(\beta, \sigma^2 \mid y) = -\frac{n}{2}\log_e(2\pi\sigma^2) - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}$$

这被称为对数似然函数且通常以 ℓ 表示。因为 ℓ 视观察值 y_i 为固定的且参数是变化的，因此我们可以写成 $\ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$，以表示 ℓ 是回归参数的一个多变量函数。

我们要寻找最大似然参数估计值，即最大化 $\ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$ 的 β_0，$\beta_1, \cdots, \beta_p$，以 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ 表示。我们可以从最小化 $\sum_{i=1}^{n}[y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})]^2$ 的可能性得到最大化 $\ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$ 的 β_j。为了找出最大似然估计值 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$，我们必须对每一个参数的对数似然 ℓ 取偏微分，令偏导数为 0，且同时解具有 $(p+1)$ 个方程的 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ 系统。在正态分布因变量的例子里，这些估计方程与那些运用最小二乘法的估计方程相同。然而，这一点并不适用于其他广义线性模型。

对 $(\beta_0, \beta_1, \cdots, \beta_p)$ 偏微分 $\ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$，以 $\frac{\partial \ell(\beta)}{\partial \beta}$ 表示，并令偏导数为 0，使得 $(X'X)\beta - X'y = 0$。因此，$\hat{\beta} = (X'X)^{-1}X'y$ 基本上为 y 的线性转换。图 5.1 画出了从一简单回归模型 $y = \beta_0 + \beta_1 X_1 + e$ 得出的 β_0 和 β_1 值的函数的对数似然表面。β_0 和 β_1 的最大似然估计值就是使对数似然函数达到最大值(两正切线交点)时的 β_0 和 β_1 值。图 5.2 通过将 β_1 固定于其最大似然值 $\hat{\beta}_{1_{(ML)}}$，显示了 β_0 的对数似然函数单一面向图。注意，对数似然函数达到最大值的正切线为一斜率为 0 的水平线，并且，它所对应的 β_0 值即为其最大似然估计。图 5.3 通过将 β_0 固定于其最大似然值 $\hat{\beta}_{0_{(ML)}}$，显示了 β_1 的对数似然函数单一面向图。

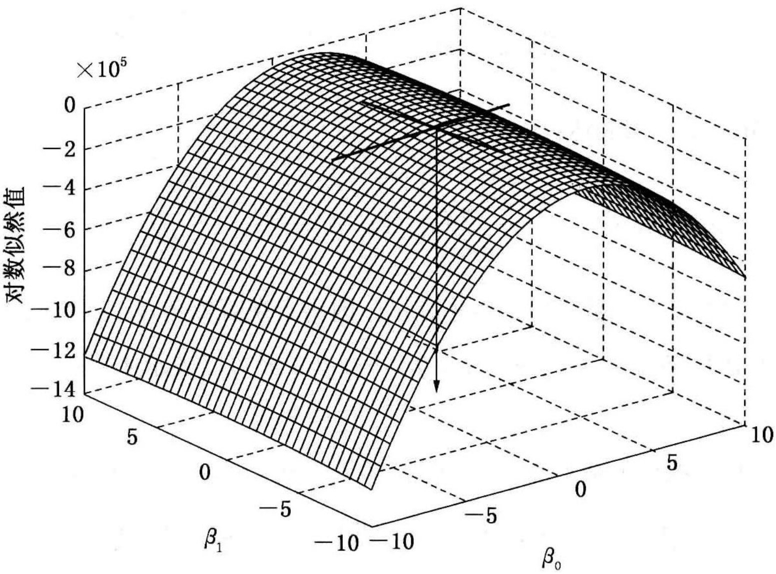


图 5.1 β_0 和 β_1 对数似然值的反应表面

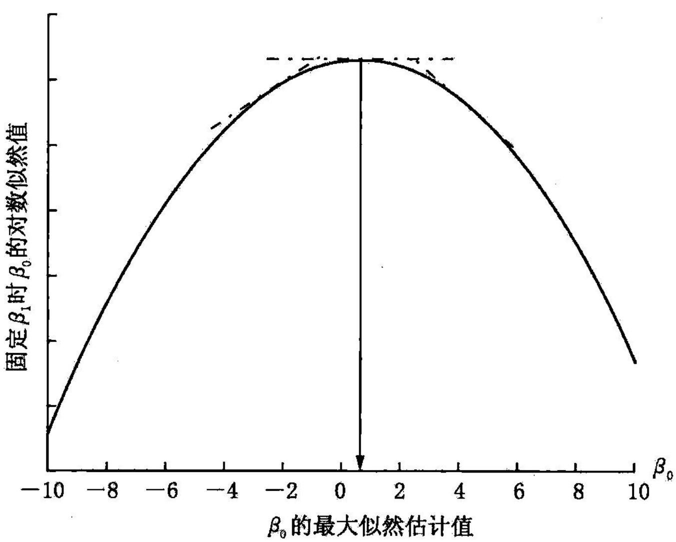


图 5.2 固定 β_1 为其最大似然值时，单一面向的 β_0 对数似然

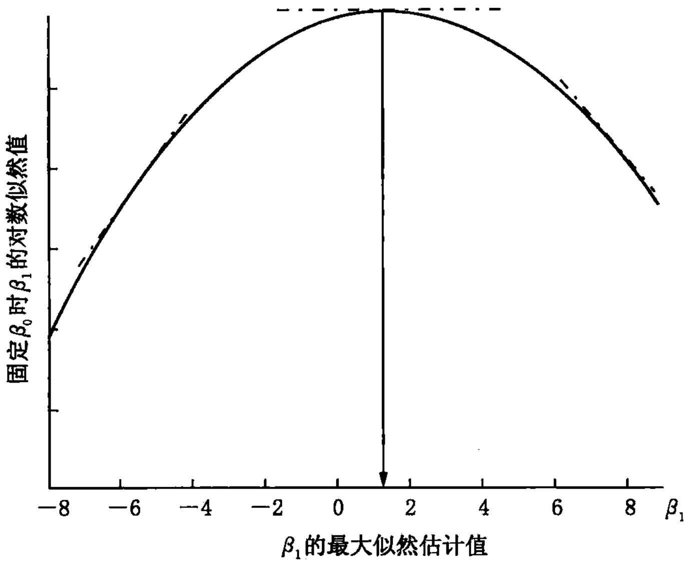


图 5.3　固定β_0为其最大似然值下，单一面向的β_1对数似然

通常在任何β_1值的正切线梯度可通过计算对数似然函数的一阶导数来决定。

图 5.4 显示了邻近 $\hat{\beta}$ 的一阶导数(正切线的斜率)的值。注意，在 $\hat{\beta}_{1(ML)}$ 处的导数为 0，但在其他β_1值时并不为 0。

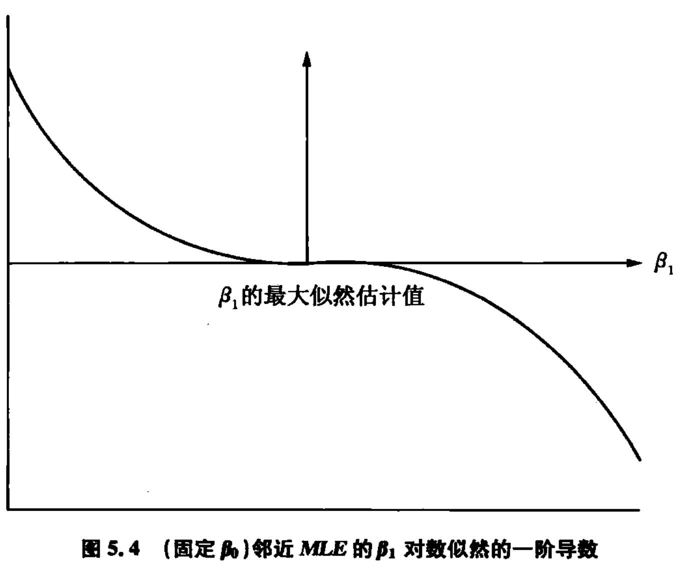


图 5.4　(固定β_0)邻近 *MLE* 的β_1对数似然的一阶导数

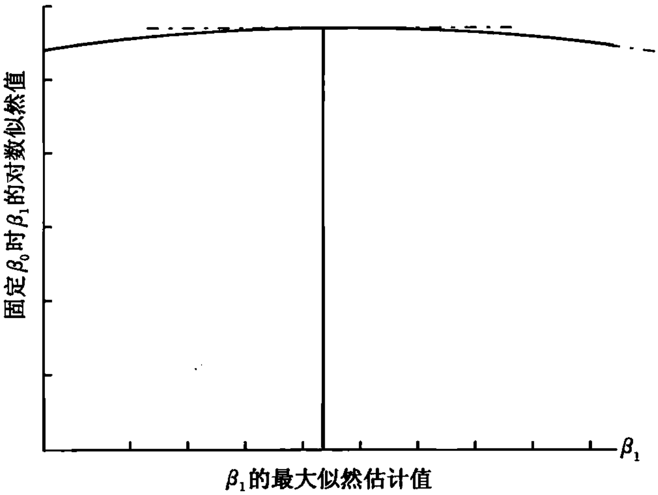


图 5.5　一个不精确的 β_1 估计值的似然方程

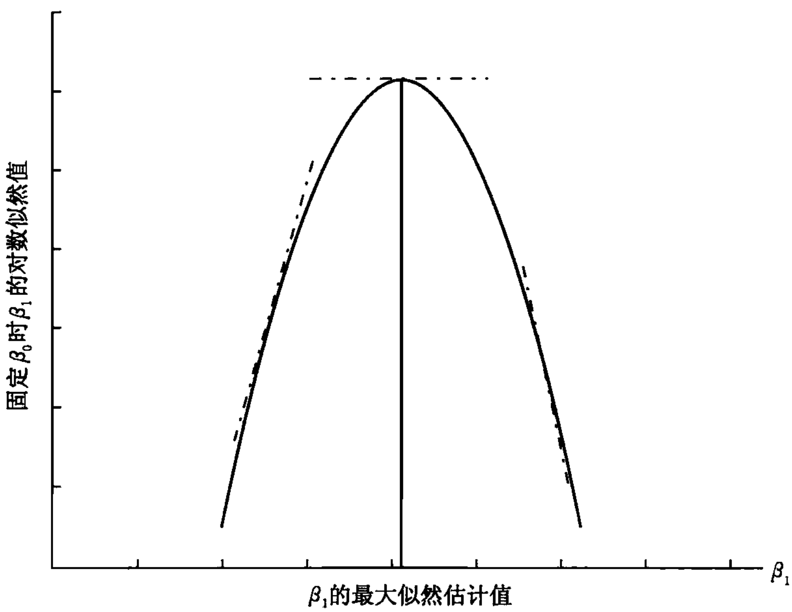


图 5.6　一个精确的 β_1 估计值的似然方程

让我们来检视两个不同的似然函数，并决定关于估计值 $\hat{\beta}_{1(ML)}$ 的精确度。图 5.5 显示了一个似然函数(根据由 10 个个案样本的回归模型产生的数据组)，在邻近 $\hat{\beta}_{1(ML)}$ 的大部分区域，平坦且宽广。换言之，该广大区域中的 $\hat{\beta}_1$ 几乎都产生相同的似然值，且我们不能精确地决定最好的 β_1 估计值。在此区域中，$\hat{\beta}_1$ 的正切线的斜率都接近于 0。相反，在图 5.6 中，另一个似然函数(根据由 1000 个个案样本的同上述回归模型产生的数据组)则更为“尖锐”，且 β_1 的最大似然估计值能更精确地被决定。如图 5.6 所示，在 β_1 的最大似然估计值 $\hat{\beta}_{1(ML)}$ 附近的正切线斜率，在 β_1 值远离 $\hat{\beta}_{1(ML)}$ 值之后突然变得陡峭了。

非正态回归的广义线性模型会产生一个无法用分析方法来估计回归参数的方程系统。我们必须求助于通过重复循环的计数法来最终逼近最大似然估计值的许多种迭代数值运算法。

第 6 章 | 离差和拟合优度

模型化的目标是要找到一组自变量，以便使模型中的 μ_i 对于观察值 y_i 有较好的拟合。对于正态分布的因变量而言，一个拟合优度的准则为 $\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$，$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$ 越小，模型 $\hat{\mu}_i=\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}$ 就越拟合数据。这就是误差平方和在最小二乘中或最大似然估计的回归参数 β_j 中的最小化。当用误差平方和除以总体方差 σ^2 时，就得到了离差，它是一个根据拟合优度的统计准则，解释广义线性模型在拟合数据方面的优势的项目。

其他广义线性模型（如泊松和 logistic 回归模型）的离差，都不同于正态的例子，并且彼此之间也各不相同。因为，从这些模型中产生的数据具有各自独特的概率函数或似然函数。离差也可以被用来比较两个模型的拟合，这是相减（两个）离差后实现的。例如，如果一个包含较多回归参数的模型的离差与一个包含较少回归参数的模型的离差相比并没有小很多，我们就可以选择包含较少回归参数的那个模型，因为它更为简洁且更容易解释。

较复杂模型（以下称完整模型）与丢弃某些参数的较简单模型（以下称简化模型）的离差间的差异，也可被用来检验虚无假设，即完整模型中的额外参数都等于零。该差值为卡方分布统计量（在虚无假设下，额外的参数都等于零），其自由度等于完整和简化模型两者间参数数目的差异。也就是说，对于这些广义线性模型，完整和简化模型的拟合差异可以被直接用于统计检验中，因为其值在虚无假设下，即额外的参数都等于零，为卡方分布。自由度等于参数数目间差值。然而，常态模型在其离差中有一个麻烦（nuiance）参数，即未知的总体方差 σ^2。如果 σ^2 是未知的且需经由样本估计，我们就无法使用卡方分布而必须

使用F分布。

现在让我们正式定义离差。离差(D)的定义为 $2[\ell(b_{max} \mid y)-\ell(b \mid y)]$，其中 $\ell(b_{max} \mid y)$ 是对数似然，当对于每一个观察值的估计的条件平均数 $\hat{\mu}_i$ 都被设为其被观察到的值 y_i。这个模型会产生最大的对数似然，因为对于每一个样本成员都有一个个别的参数 $\hat{u}_i = y_i$。参数同样本成员一样多；因此，模型会完全地拟合数据。我们使用类似这样的一个模型并没有简化任何事情。然而，它的确显示出被观察到的样本 y_i 的最大似然。它为我们评价其他不太复杂的模型提供了一个基准。方括号中的第二项 $\ell(b|y)$ 是较简单模型的对数似然，其参数为 $\beta=(\beta_0, \beta_1, \cdots, \beta_p)'$，通过最大化似然估计并且用来产生 $\hat{\mu}_i$。对数似然差值的两倍，给了我们关于此模型与数据拟合程度的一个提示。

例如，如果我们有10个 y_i 的观察值，y_i 为一特别的正态分布因变量，且有三个自变量 X_{i1}，X_{i2} 和 X_{i3}，通过替代 y_i 于对数似然的 $\hat{\mu}_i$，则 $\ell(b_{max}|y)$ 可以被计算出来，而 $\ell(b|y)$ 可以通过估计 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 及以 $\hat{\mu}_i=\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\hat{\beta}_3X_{i3}$ 替代在对数似然方程中100个估计的 $\hat{\mu}_i$ 中的任何一个计算出来。最大的模型包含100个参数，而(目前)考虑的模型仅包含四个参数。离差 D 表示较简单模型相对于完全拟合数据的最大模型的优势。

为了比较这两个模型，我们计算它们各自的离差，并用较简单或简化模型的离差(D_R)减去较复杂模型或完整模型的离差(D_F)。亦即，我们计算 D_R-D_F。如果差值很大，则证据显示完整或较复杂模型对数据的拟合更好。简化或较简单模型的离差永远会比完整或较复杂模型的离差大。关键是它能否足够大，以确保在模型中增加额外的回归参数。

使用离差进行假设检验

我们可以通过在一正式统计检验中的离差间的差额，假设检验在完整模型中额外的参数都等于0。在虚无假设中，额外的参数都等于0，(D_R-D_F) 指自由度等于两模型中参数数目差额，即完整模型中额外参数数目的卡方分布。对于广义线性模型，如泊松和 logistic 模型，我们可以计算 (D_R-D_F) 的数值，且在

一个相应的自由度(也就是参数数目的差值)下查询卡方表,以决定该卡方是否显著。

拟合优度

我们可以通过比较相关的离差来比较各种其他备选模型的拟合。我们也可以比较各种连结函数的拟合。此外,我们应该检验模型的残差以确定是否有些观察值有大的残差,如果有,则表示就这些观察值而言,模型对自变量的拟合不好。

如果最终选择模型的离差很大且查询相应自由度(样本数目减去参数数目)的卡方分布时为统计显著,则可能表示过度离散(overdispersion)。过度离散意味着因变量对于其条件平均数估计的方差较用于模型中的概率分布所期望的大。例如,泊松分布假设分布的方差等于其平均数。如果数据显示它较大,则回归参数的标准误可能需要向上调整,即使是无偏参数估计值。

通过残差分析衡量拟合优度

同基于正态分布的回归模型一样,适合广义线性模型的各种残差分析可以被用来指明没有被模型解释清楚的拟合不足的观察值。可以用残差图来检验因变量的值对模型的预测值发生系统性偏移的情况。残差分析可以帮助再次指明一个能显示更拟合数据的模型。例如,它可以提供需要二次项的证据,如一个或多个自变量的平方项,或者,它也可以提供证据显示因变量的观察值是相关的,并因此违反了似然函数的独立假设。有些广义线性模型可以适用于观察值之间具有相关性的情况,该相关性源于对相同个体的重复测量结果或同构型群体内的个体聚集效应,如学校中的班级。然而此处我们不讨论这些较为复杂的模型。

对广义线性模型衡量拟合优度的最简单的残差为 Pearson 残差,也就是:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{var}(\hat{\mu}_i)}}$$

帽子数值表示一个观察值的平均数和方差都是从模型估计而来的预测值。例如,在 logistic 回归模型的例子里，$y_i = 1$ 或 0，

$$\hat{\mu}_i = \hat{\pi}_i = \frac{e^{\sum_{j=0}^{p} \hat{\beta}_j X_{ij}}}{1 + e^{\sum_{j=0}^{p} \hat{\beta}_j X_{ij}}}$$

且 $\mathrm{var}(\hat{\mu}_i) = \hat{\pi}_i(1 - \hat{\pi}_i)$。注意,与正态模型将各观察值之间的方差 σ^2 假设为固定有所不同,在 logistic 回归中,每一个观察值 y_i 都有根据回归参数的一个独特方差以及相对应的自变量 X_i 的值。logistic 回归模型的 Pearson 残差为：

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

大的 r_i 值表示对于该观察值失拟。

在泊松回归模型的例子中：

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{var}(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

其中：

$$\hat{\mu}_i = \hat{\lambda}_i = e^{\sum_{j=0}^{p} \hat{\beta}_j X_{ij}}$$

对于泊松分布,方差等于其平均数,因此在前述方程中，$\hat{\mu}_i = \mathrm{var}(\hat{\mu}_i)$。对于泊松模型,一个计数的结果变量 y_i 的方差会随着该观察值的期望平均数而变化。

另一个常被使用的残差是离差残差。离差由个别观察值的离差组成。每个单一观察值 y_i 对于离差的贡献,为该模型对于个别观察值 y_i 拟合度的一个测量。如 Pearson 残差,它的定义是根据关于某一特定广义线性模型的离差的形式而定的。

离差残差为：

$$r_i = sign(y_i - \hat{\mu}_i)\sqrt{2(\ell_i(y_i) - \ell_i(\hat{\mu}_i))}$$

$r_i = sign(y_i - \hat{\mu}_i)$ 项表示残差$(y_i - \hat{\mu}_i)$是否为正或负。当第 i 个人的条件式分布平均数为该个人实际的因变量得分时，$\ell(y_i)$项为对数似然的值；$\ell(\hat{\mu}_i)$为

对数似然。当由该模型产生的条件平均数替代到对数似然中时，根号下的项为第 i 个观察值对于整体离差的贡献，如前文所示，它等于 $2\sum_{i=1}^{n}[\ell(y_i)-\ell(\hat{\mu}_i)]$。例如，泊松分布的离差残差为：

$$r_i = sign(y_i-\hat{\mu}_i)\sqrt{2[y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)]}$$

第 7 章 | logistic 回归

我们在第 4 章中简单介绍过 logistic 回归。运用前面讨论过的一些模型化的概念，我们现在要显示如何将 logistic 回归整合到拟合广义线性模型的架构中。一个在真实数据中运用 logistic 回归的例子也会被呈现。贝努利分布(Bernoulli distribution)为二项式分布的特例，用来模型化 0—1 二元结果或因变量，如治疗成功(1)和治疗失败(0)，或烟瘾复发(1)和没有复发(0)。事件的哪一个方面被编码为 1 或 0 完全是随意的。一般而言，关注的事件被编码为 1。其他二元结果变量的例子有死亡 vs. 生存和犯罪 vs. 没犯罪。社会科学中其他学科的许多结果也是二元的。它们都是 logistic 回归模型的因变量。

贝努利分布或二元分布为 $f(y \mid \pi) = \pi^y (1-\pi)^{1-y}$，其中 π 为 $y=1$ 时成功的结果的概率。(y 的值)有两个结果：如果结果是成功的，则 $y=1$；如果结果是失败的，则 $y=0$。注意，如果我们将 $y=1$ 替代至概率分布中，就会得到 $f(1 \mid \pi) = \pi$。如果我们替代 $y=0$，则得到 $f(0 \mid \pi) = 1-\pi$。如果只有两个结果，且其中一个(结果)的发生概率为 π，则另一个结果的发生概率必然为 $1-\pi$，因为(两)概率的总和必然等于 1。

这是一个来自于指数家族的相当简单的概率分布，可以很容易地以指数形式表示：

$$f(y \mid \pi) = e^{y\log_e\left(\frac{\pi}{1-\pi}\right)+\log_e(1-\pi)} = e^{y\theta-\log_e(1+e^{\theta})}$$

其中，标准参数为 $\theta = \log_e \dfrac{\pi}{1-\pi}$ 且 $b(\theta) = \log_e(1+e^{\theta})$。在这个例子中，离散参数 $\phi = 1$ 且 $c(y, \phi) = 0$。标准连结函数为 $\theta = \log_e \dfrac{\pi}{1-\pi}$，且 θ 是我们以协变量

的一个线性函数所模型化的参数——即 $\theta=\beta_0+\sum_{j=1}^{p}\beta_j X_j$。方差函数为 $b''(\theta)=\pi(1-\pi)$。它是一个单一参数分布，且方差 $\pi(1-\pi)$ 与其平均数 π 有关。贝努利分布的对数似然为 $\ell(\pi\mid y)=y\log_e\pi+(1-y)\log_e(1-\pi)$。全样本的对数似然为 $\ell(\pi_1,\cdots,\pi_n\mid y_1,\cdots,y_n)=\sum_{i=1}^{n}[y_i\log_e\pi_i+(1-y_i)\log_e(1-\pi_i)]$。

与前文相同，我们假设每个观察值或样本成员来自一个有其独特参数 π_i 的贝努利分布。我们想把对数似然 $\ell(\pi|y)$ 写成一个回归参数（即 $\ell(\beta|y)$）的函数，因为它们而不是 π_i，才是我们想要从数据中估计的。代数运算可以显示为：

$$\pi_i=\frac{e^{\sum_{j=0}^{p}\beta_j X_{ij}}}{1+e^{\sum_{j=0}^{p}\beta_j X_{ij}}}$$

而这便是我们替代至对数似然的部分。

为了找到参数的最大似然估计值，我们以回归参数的一个函数的形式替代 $\pi_i(\beta)$ 至对数似然方程，并对于每一个参数偏微分令偏导数为 0，并求取回归参数向量的解 β。因为这些方程的参数为非线性的且不能以分析方法解出，因此必须用迭代运算，例如运用反复重加权（iterative reweighted）最小二乘法来求解 β。

logistic 模型的离差为 $D=2[\ell(y\mid y)-\ell(\hat{\beta}\mid y)]$，其中 y 为二元结果变量 $(y_1,y_2,\cdots,y_n)'$ 的向量，且 $\hat{\beta}$ 为最大似然回归参数估计值的向量。似然饱和（saturated）模型为 $\ell(y|y)$，其中参数 π_i 的最大似然估计值为 y_i；即对于每一个观察值都有一个独特的参数估计值。对于贝努利分布，$\ell(y\mid y)=\sum_{i=1}^{n}[y_i\log_e y_i+(1-y_i)\log_e(1-y_i)]$。注意，当 y_i 是 0 或 1 时，方括号中的项目等于 0，所以 $\ell(y\mid y)=0$。因此，离差可简化为：

$$D=-2\ell(\hat{\beta}\mid y)=-2\sum_{i=1}^{n}[y_i\log_e\hat{\pi}_i(\hat{\beta})+(1-y_i)\log_e(1-\hat{\pi}_i(\hat{\beta}))]$$

其中 $\hat{\pi}_i(\hat{\beta})$ 表示最大似然估计值，$\hat{\pi}_i$ 为最大似然回归参数估计值 $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)'$ 的一个函数。同前述：

$$\hat{\pi}_i(\hat{\beta}) = \hat{p}_i = \frac{e^{\sum_{j=0}^{p}\hat{\beta}_j X_{ij}}}{1 + e^{\sum_{j=0}^{p}\hat{\beta}_j X_{ij}}}$$

因此，对于一个特殊的 logistic 回归模型，$D = -2\ell(\hat{\beta} \mid y)$ 测量了设想中的模型的拟合数据程度。如前述，两个不同回归模型(其中一个是简化模型，包含的参数是完整模型中参数的一个子集)离差间的差异，就是具有自由度等于完整和简化模型两者参数数目差值的卡方分布，可以用来比较一个特定的模型与一个参数较少的简化模型的相对拟合。如果服从卡方分布的$(D_R - D_F)$在适当的自由度下不显著，则证据显示完整模型中的多余变量是不必要的，一个有较少回归参数的较简单模型就已足够。

如果我们想要检验特定的单个回归参数是否在统计上显著，则可以构建 $\frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$ 这一 t 比率，并将该 t 比率与一个自由度为 $(n-p-1)$ 的 t 分布相比较，以确定它是否统计显著。(该比率式中的)分母为回归参数估计值的标准误。我们也可以通过取在模型中有该参数的离差差额和无该参数的离差差额来检验回归参数的显著与否，并且查自由度为 1 的卡方表。

如同所有的回归模型一样，logistic 回归的主要兴趣是在参数估计值、估计标准误、t 比率以及统计显著程度。这些信息构成了所有估计广义线性模型的统计软件的核心产物。对于每种广义线性模型，有关其对于因变量的效应的回归参数都有不同的解释。

对于 logistic 回归，连结函数为其平均值的非线性函数，也就是说，$\log_e\left(\frac{\pi_i}{1-\pi_i}\right)$ 为回归参数的一个线性函数。因此，β_j 表示因 X_{ij} 增加一个单位时在 $\log_e\left(\frac{\pi_i}{1-\pi_i}\right)$ 中的改变，而并不是指 π_i(贝努利分布的平均数)的变化。

解释 β_j 是非常困难的，因为它反映了 $\log_e\left(\frac{\pi_i}{1-\pi_i}\right)$ 的变化。如果我们指数化 β_j(即 e^{β_j})，则 e^{β_j} 测量了 X_i 每增加一个单位时发生比率的变化。让我们以一个简单的、只有单一自变量的 logistic 回归的例子来说明。令 $\log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1$。图 7.1 显示了此函数。

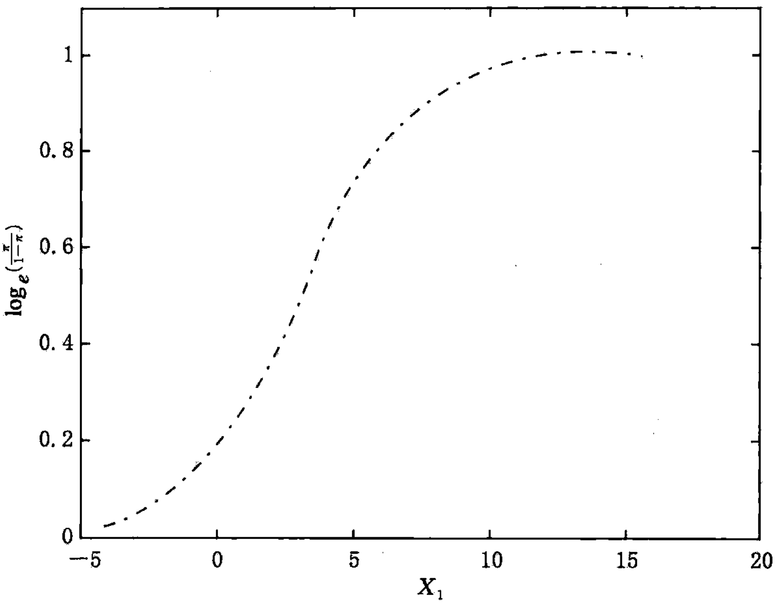


图 7.1 logistic 回归函数

如果我们将此 logistic 回归方程的两边指数化，则：

$$\frac{\pi}{1-\pi}=e^{\beta_0+\beta_1 X_1}=e^{\beta_0}e^{\beta_1 X_1}$$

如果我们将 X_1 增加一个单位，则：

$$\frac{\pi^*}{1-\pi^*}=e^{\beta_0+\beta_1(X_1+1)}=e^{\beta_0}e^{\beta_1 X_1}e^{\beta_1}$$

我们可以通过将第二个方程除以第一个方程得到 e^{β_1}。即：

$$\frac{\dfrac{\pi^*}{1-\pi^*}}{\dfrac{\pi}{1-\pi}}=\frac{e^{\beta_0}e^{\beta_1 X_1}e^{\beta_1}}{e^{\beta_0}e^{\beta_1 X_1}}=e^{\beta_1}$$

因此，e^{β_1} 代表因 X_1 增加一个单位而导致的发生比的增加或减少。如果 β_1 为正，则 e^{β_1} 大于 1，表示因 X_1 增加一个单位而导致的发生比的增加。例如，如果 β_1 等于 0.1，则 $e^{0.1}$ 等于 1.11，X_1 增加一个单位导致了发生比增加 11%。一个

大的β_1正值会导致大的e^{β_1}值，也因此代表因X_1增加一个单位而导致的发生比的大幅增加。然而，如果β_1为负，则e^{β_1}小于1，代表因X_1增加一个单位而导致的发生比的减少。例如，如果β_1等于−0.1，则$e^{-0.1}$等于0.90，X_1增加一个单位导致了发生比减少10%。一个大的β_1负值会导致小的e^{β_1}值及发生比的大幅减少。

除了logistic连结外，还有其他可以用来模型化二元变量的连结函数（如probit），但它们基本上都会给出与logistic回归相同的结果。因为logistic的参数较其他模型的参数更易于解释，对二元因变量而言，它是被最广泛使用的模型。

logistic回归实例

本书第一作者参与了以州级社会指标来估计药物使用流行性及治疗需求的研究。采用州级社会指标方法的基本假设为：一州内的各县社会、人口及经济特性或其他在当地的计划实体（planning entities），都与药物使用流行性及治疗需求有关。县级层次的社会指标实例可从各县、州及联邦机构获得，如家庭收入的中位数；依年龄、性别及种族划分的人口分布；与酒精或药物相关的意外事件率；暴力犯罪率。

虽然直接在州的全部人口中调查药物与酒精的使用可能是获取信息的最佳方法，但此方法有一些致命的缺点。衡量药物使用问题的逐年调查费用很高。即使在州层级也有4000或5000名受访者这样相对较大的样本数，对估计县级的药物使用来说它仍然太小。一个州通常不会有超过50甚或100个县，因此，平均的县级样本数会在50至100的范围内。这些样本数对于推论任何县级层次上关于药物使用及治疗的需求来讲实在是太小了。

如果我们孤立那些根据州级电话访问数据得来的、有助于预测县级药物使用和治疗测量的县级社会指标变量，就可以达成两个重要的目标。第一，可以用logistic回归模型对县级层次产生以模型为基础的估计值，这会比仅仅根据

每一个县级小样本得到的调查估计值更为精确。也就是说，对于县级层次的药物使用与流行程度的直接调查估计值而言，是根据 50 至 100 个受访者的信息得到的，而根据 logistic 模型得到的间接估计值则是使用一州内所有 4000 或 5000 名受访者的数据得到的。一旦该模型被估计，一个对某种药物或治疗需求的县级的普遍性测量，就可以通过将每个县级社会指标值代入 logistic 回归方程来获取关于该县的药物使用情况或治疗需求概率。在这种情况下，我们不是对估计药物使用的发生比感兴趣，而是对药物使用的概率感兴趣。因为：

$$\frac{\pi}{1-\pi}=e^{\sum_{j=0}^{p}\beta_j X_j}$$

稍做代数转换就会有：

$$\pi=\frac{e^{\sum_{j=0}^{p}\beta_j X_j}}{1+e^{\sum_{j=0}^{p}\beta_j X_j}}=\frac{1}{1+e^{-\sum_{j=0}^{p}\beta_j X_j}}$$

因此，只要 β_j 被估计出来，我们就可以将某个县的相关社会指标值 X_j 代入前述公式，以获得该县的药物使用期望值。

而且，当昂贵的药物使用调查数据无法获取时，我们还可以用这个模型来预测未来的药物使用情况。许多县级的社会指标是动态变量，其值会随时间改变（如各种犯罪率）。因此，社会指标数据可在未来数年被收集，并被代入前述方程，以预测未来几年的药物使用情况。在大多数例子中，仅有少量的社会指标变量对于药物使用流行性来说是重要的预测因素，因此只需要收集少量的信息代入前述方程中。logistic 回归方程，也就是前述方程，捕捉了县级社会指标与县级药物使用概率之间的关系。当社会指标随时间改变时，对药物使用及流行程度的预测也会有相应的改变。

这里呈现的 logistic 回归模型是依据一个在南达科他州（South Dakota）进行的调查研究。用来校准 logistic 回归的社会指标模型（也就是因变量为受访者的药物使用二元变量）的药物使用和治疗需求的电话调查，是一个总数为 4205 名受访者的样本。南达科他州有 66 个县，且各县级的样本数为 4 到 896。正如所预期的那样，县级样本数与县的人口规模大约是成比例的。县级样本数

大多为 50 或更小。若干对药物使用流行性的测量被模型化，包括去年饮酒过度、去年违法药物使用、酒精治疗需求及药物治疗需求。41 个县级层次的社会指标变量数据通过各种渠道搜集而来。它们包括酒精和药物使用指标（如因使用或持有药物的成人逮捕率）、社区解组（disorganization）（如离婚率）、小区犯罪和暴力（如因暴力犯罪的成人逮捕率）、人口特性（如该县的白人人口比例）、社会经济剥夺（如失业率）、酒精及药物的可获得性（如到达最近州际高速公路的距离）、学业失败或缺乏承担义务（如高中辍学率），以及与物质滥用间接相关的社会问题（如青少年生育率）。

由于指标很多且彼此高度相关，故需要使用因子分析来将变量分组，并选一些样本代表每一群变量，或者说来测量每一个因素，这会减少最初的那组社会指标数目。如果减少的社会指标组（数目）同时作为自变量在 logistic 模型中预测各种药物使用测量仍太多，就需逐步用 logistic 回归模型来建立更加简约的模型。

我们对该例子使用这些模型中的其中一种，因变量是二元的，指酒精使用或药物使用的某种干预或治疗需求。如果该受访者有干预需求则被编码为 1，否则为 0。四个县级的社会指标或自变量为青少年违法饮酒（JLLV），以每 1000 个青少年违法饮酒数目测量；青少年犯罪被捕率（JVC），以每 1000 个青少年被捕率测量；年轻男子比例（YM），测量单位为占该县人口的比例；该县收入中位数（MI），测量单位为元。logistic 回归分析结果于表 7.1 中呈现。

表 7.1　药物使用干预之 logistic 回归模型

变　量	参数估计值	标准误	t 比率	显著水平	发生比率
截距	3.7757	0.4202	8.985	<0.0001	—
JLLV	−0.00707	0.00317	−2.230	0.0321	0.993
JVC	0.1964	0.0586	3.352	0.0019	1.217
YM	−0.0471	0.0153	−3.078	0.0040	0.954
MI	−0.00007	0.000016	−4.375	0.0001	0.9999

表 7.1 中的 logistic 回归分析结果为 SAS 输出结果的修正。注意 *JLLV* 和 *MI* 的估计回归参数相当小。它们各自有 0.993 和 0.999 的发生比。部分原因

与自变量的测量尺度有关。青少年违法饮酒的发生比率意味着在1000名青少年中每增加一起违法饮酒，就会使药物治疗干预需求的发生比降低0.7%($= 100(1-0.993)\%$)，而中位数收入每增加1元，就会使干预需求的发生比降低0.1%($= 100(1-0.999)\%$)。

青少年违法饮酒的标准差为每1000个青少年中有14.1起违法，因此，每1000个青少年中增加一起违法是相对较小的改变。如果我们改变*JLLV*的测量尺度为标准差单位，则*JLLV*增加一个标准差所导致的发生比率的变化为$(0.993)^{14.1}$或0.906。因此，*JLLV*的标准差每增加一个单位，会使干预需求的发生比降低9.4%。即使就标准差单位而言，这对于*JLLV*也只是一个中等程度的效应。

对于中位数收入来说，更加有意义的变化可以用\$1000单位而非\$1单位来测量。南达科他州66个县的中位数收入位于\$11502到\$34286之间。如果我们将中位数收入尺度改为以\$1000为单位，则相关的发生比率变成了0.368。这是一个大的效应，因为中位数收入每增加一个\$1000，就会使干预需求的发生比降低63.2%。

每1000名青少年的暴力犯罪被捕率为0至3.73，其中许多县显示没有逮捕，因此原本的每1000人中有1人被捕的度量标准(metric)设定似乎是合理的。关于*JVC*的发生比率为1.217，代表每1000人中一次单一逮捕会使干预需求的发生比增加21.7%，为中度效应。年轻男子(15—34岁)于各县中所占的比例为9%到23%之间，且标准差为2.7%。年轻男子的比例每增加一个标准差，对干预需求的发生比的影响效应为$(0.954)^{2.7}=0.881$，即年轻男子的比例每增加一个标准差会使干预需求的发生比降低11.9%，这是一个中等程度的效应。

青少年犯罪对干预需求发生比的正效应和中位数收入的负效应合乎预期。年轻男子比例的负效应和青少年违法饮酒率的负效应则有些违反我们的直觉经验。或许高违法饮酒率代表相关法律执行单位在检查低龄饮酒时的高度谨慎，因此减少了干预需求。然而，必须谨记，这四个社会指标的效应都是在调整了其余三个变量的影响效应之后的结果。在模型中各变量间的相互关系的形

态可能会造成调整过的效应(如男性百分数)与基于二元模型的未调整过的效应的方向相反的情形。

为了获取某县的干预需求预测值,我们只需简单地将它所对应的四个社会指标的值代入下述公式:

$$\hat{p}_{need} = \frac{1}{1+e^{-(3.7757-0.00707JLLV+0/1964JVC-0.0471YM-0.00007MI)}}$$

第 8 章 | 泊松回归

泊松回归模型假设回归模型的随机成分有一个特殊的概率分布，即泊松分布。泊松分布适用于计数数据。所谓计数数据，指在给定的一段时间内，一个特定事件发生的次数。下面的计数数据实例适用于泊松分布：在给定的一段时间内(如一年)，某个繁忙的交叉路口所发生的交通意外事件数；接线总机在一小时内所接到的电话来电数目；一年内一个罪犯的犯案次数；五年内一个药物成瘾者的治疗次数；以及在给定的一段时间内，某一特定医院因药剂过量者的急诊室进入许可数目。

泊松概率密度函数的表达较正态概率密度函数简单得多。

$$f(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

其中 $e = 2.7183$（自然对数的底），且 $y! = y(y-1)(y-2)\cdots 1$。例如，$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$。由此可见，泊松分布中仅有一个参数 λ，它是在一给定时间内的事件平均数。图 8.1 表示在不同 λ 值下的泊松分布。

泊松分布为指数家族的成员，因为：

$$f(y \mid \lambda) = e^{(y\log\lambda - \lambda - \log y!)} = e^{(y\theta - e^{\theta} - \log y!)}$$

因此，标准参数 θ 等于 $\log_e\lambda$，它同样也是标准连结。$b(\theta)$ 的方差函数为 e^{θ}，其二次导数为 $b''(\theta) = e^{\theta}$。因为 $e^{\theta} = \lambda$，泊松分布的方差等于其平均数。当泊松分布的平均数增加，其方差也会增加。随机变量 Y 只能为非负数的整数值——也就是 0，1，2，…。泊松分布是左偏的。

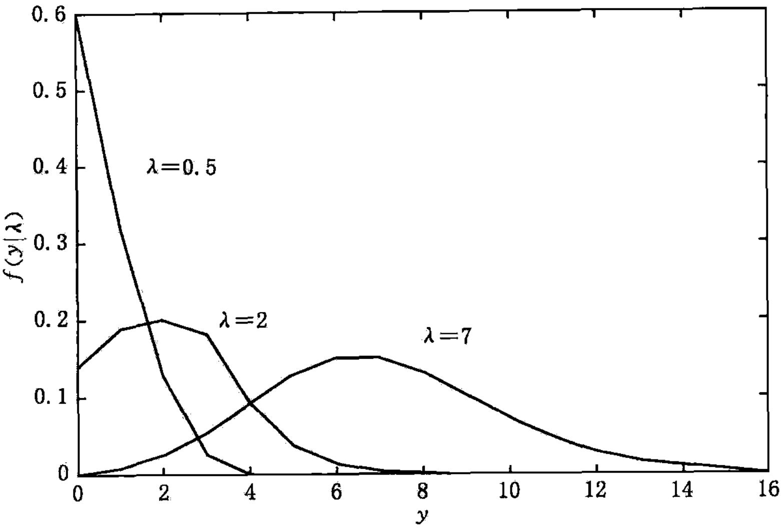


图 8.1　在不同 λ 值(0.5, 2 和 7)下的泊松分布

对于假设有一个标准连结的泊松回归模型而言，我们假设 $\log_e\lambda$ 为线性模型的一个函数，即 $\log_e\lambda = \sum_{j=0}^{p}\beta_j X_j$。我们将 $\sum_{j=0}^{p}\beta_j X_j$ 代入样本的对数似然中以求取最大化似然函数的 β_j。

因为我们关注回归模型，我们可以检验有 p 个参数的回归模型，是否比仅有一个总体平均数的简单模型(可想象成是在模型中仅有一个截距项的回归模型)更好地拟合数据。我们之前就做过对于仅有总体平均或截距参数的简化模型和有着条件期望值为 λ_i(即回归参数的一个函数)的完整模型，比较两者离差的差异。我们也可以比较完整的和简化的这两个模型，由此再比较这两个模型间离差的差额。该差异服从自由度为回归参数数目差值的卡方分布。如果卡方量为统计显著，我们就接受完整模型。如果不显著，则接受简化模型。

如前所述，离差的定义为 $2[\ell(y \mid y)-\ell(\hat{\lambda} \mid y)]$，其中 $\ell(y \mid y)$为拟合值等于数据达到的最大似然。亦即，有如同观察值一样多的参数。第二个似然 $\ell(\hat{\lambda} \mid y)$为基于预测值 $\hat{\lambda}_i$，$\hat{\lambda}_i$ 由一个 $\hat{\lambda}_i$ 为 $\hat{\beta}$ 的函数，也就是 $\hat{\lambda}_i(\hat{\beta})$的回归模型产生。对于泊松分布，离差为：

$$D = 2\left\{\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right)-(y_i-\hat{\lambda}_i)\right]\right\}$$

模型化的目标是要找到一个 D 值小的模型，因为它表示条件化平均数或期望值 $\hat{\lambda}_i$ 与观察值 y_i 相近。如果模型完全拟合数据(如同一个饱和模型)，则 D 等于 0，因为 $y_i = \hat{\lambda}_i$。我们可以从公式中看到，当 y_i 和 $\hat{\lambda}_i$ 间的差异增加时，D 会变大。因此，它可以用来衡量设想模型的拟合优度。因为 D 为个别离差项目的总和——即 $D = \sum_{i=1}^{n} d_i$ ——我们可以检验个别 d_i 并确定是否有特别大的 d_i 值。离差残差的定义为 $(y_i - \hat{\lambda}_i)\sqrt{d_i}$，其中如果 $y_i > \hat{\lambda}_i$，$\mathrm{sign}(y_i - \hat{\lambda}_i)$ 为正；如果 $y_i < \hat{\lambda}_i$，为负；如果 $y_i = \hat{\lambda}_i$，等于 0。大的离差残差可能代表失拟。

对于泊松模型，因为 $\log_e \lambda_i = \sum_{j=0}^{p} \beta_j X_{ij}$，这意味着：

$$\lambda_i = e^{\sum_{j=0}^{p} \beta_j X_{ij}} = e^{\beta_0} e^{\beta_1 X_{i1}} e^{\beta_2 X_{i2}} \cdots e^{\beta_p X_{ip}}$$

因此，在假设其他自变量不变的条件下，X_{ij} 每增加一个单位，就会通过一个 e^{β_j} 改变平均数 λ_i 的值。

泊松回归模型实例

这里叙述的泊松回归模型是一个对应于一项重要社会政策议题的实际生活实例。本书第一作者为此项目的共同研究者。该项研究涉及北卡罗来纳州囚犯的纪律违反率。监狱官及其他监狱系统的工作者观察到，囚犯的纪律违反率在近期内有较大幅度的增加。有人认为，这种增加是由于判决法的改变。在 1994 年 10 月 1 日以前，有罪的重罪犯依公平判决法(Fair Sentencing Act, FSA)判决。根据 FSA，囚犯有相对长的刑期，但可能使其减半并且有资格假释。从 1994 年 10 月 1 日起，有罪的重罪犯依结构判决法 (Structured Sentencing Act, SSA)判决。此法提供相对短的刑期且没有假释资格，即使行为良好也不能提前释放。囚犯可以通过参与某些工作或计划赚取一些时间以提早释放。然而，赚取的时间不能使一名囚犯的服刑时间少于他或她最长刑期的 83%。

因此，SSA 囚犯较之于 FSA 囚犯有更少的动机去服从监狱的纪律规范，也

因此被预期有较高的纪律违反率。对职工和囚犯安全而言，这可能会导致许多具有危害性的结果。

此研究设计旨在测定，根据 SSA 判刑的囚犯与根据 FSA 判刑的囚犯相比，是否具有显著较高的纪律违反率。此研究涉及从 1995 年 6 月 1 日起被送至北卡罗来纳州监狱服刑的所有年龄层的男女囚犯。有些囚犯(从研究中)被排除，例如那些同时根据 FSA 和 SSA 服刑的，以及因为违反假释而回到监狱的。其他的一些排除原因在此不需要讨论。在研究期间有大量的 FSA 和 SSA 囚犯，用于回归分析的数据来自北卡罗来纳州矫正部的计算机化罪犯记录。

此研究涉及分别对男性和女性模型化不同类型的纪律违反率。男性和女性之所以被分开模型化，是因为男性有较高的违反率，且预期 FSA/SSA 及其他协变量对于纪律违反率的效用在男女间有所不同，即，假设 FSA/SSA 及大多数其他协变量与性别有交互作用。我们着重用泊松回归来模型化男性囚犯的总的或整体的纪律违反率。这在此研究计划的所有泊松回归模型估计中可能是最重要的，因为监狱中的囚犯大多是男性，且他们的纪律违反率比女性囚犯高。

泊松回归模型的样本数为 11738。这些囚犯的记录没有缺失值。另外1026 名囚犯因为有缺失数据被排除在分析之外。因为依据 FSA 和 SSA 判刑的囚犯可能在重要的背景特征上有所差异，也因此被预期会与纪律违反率有关，故需要将它们作为协变量，与 FSA/SSA 这一“甄别”(treatment)或政策变量一同放入模型，以调整 FSA/SSA 政策变量对协变量的影响效应。在囚犯记录数据中有许多可能的协变量，但根据逻辑上的考虑和过去的研究，我们只选择了由某些协变量所构成的一个次集合。

主要感兴趣的变量为结构判决相对于公平判决这一政策变量，以 0—1 指标置入模型中，0 表示公平判决，1 表示结构判决。因此，公平判决代表参照组，用来和结构判决比较。此 0—1 变量的回归参数代表了结构判决相对于公平判决的影响效应。在此模型中，该参数已经调整了其他协变量的影响效应。

用一个三级的类别变量反映囚犯被判的犯罪类型。第一级是暴力犯罪；第二级是财产犯罪；第三级是公共秩序犯罪。我们假定因暴力犯罪而被判刑的囚犯会较其他两种非暴力犯罪类型的囚犯有更高的纪律违反率。因为有三种犯

罪类别，需要两个 0—1 指标变量来代表这个三级的犯罪变量。注意，第三个指标是多余的，因为只要知道任何两个指标变量的值，剩余的一个指标变量值就可以被确定了。因为三个指标变量之间具有完全的多重共线性，所以其中一个必须从模型中去除。我们采用舍弃代表公共秩序犯罪类别的 0—1 指标变量。两个剩余的指标变量被编码如下：暴力犯罪编码为 1，如果该囚犯为暴力犯罪群体中的一员，为 0 则否；财产犯罪编码为 1，如果该囚犯为财产犯罪群体中的一员，为 0 则否。同样的，关于财产犯罪类的回归参数测量的是作为财产犯罪群体中的一员相对于作为公共秩序犯罪群体中的一员的效应。

种族有三个类别：黑人、白人和其他。白人被选为参照组。模型中包含一个 0—1 的黑人指标变量和一个 0—1 的其他种族的指标变量。因此，黑人的回归参数反映了相对于白人而言，黑人的纪律违反率。同样的，其他种族群体指标变量反映了相对于白人群体而言，属于其他种族群体的纪律违反率。

有一个三类别变量代表先前的监狱经验与纪律违反的组合。这些类别为：先前的监禁和至少一次（纪律）违反、先前的监禁和没有（纪律）违反、没有先前的监禁。第三个类别，没有先前的监禁为参照组。前两个监狱经验变量都以适当的指标变量来定义。

有一个 0—1 的指标变量代表该囚犯对于目前的监禁是否在被逮捕时为缓刑。获缓刑的囚犯被编码为 1，而未获缓刑者被编码为 0。因此，这个变量代表相对于没有缓刑，缓刑对违反率的影响效应。一个指标变量被构造来表示酒精依赖，1 表示对于酒精依赖有高度风险，0 代表没有高度风险。另有一个相似的变量被用来表示药物依赖。

有四个连续性自变量：开始服刑时的年龄，以年表示；在研究之前就已经在北卡罗来纳州监狱服刑的时间，以年表示；适用目前刑期的（jail credit）年数；预期的刑期长度，以年表示。对于这些连续性变量，相关的回归参数反映了自变量每增加一个单位尺度（如年）所导致的纪律违反率的改变。

对于该项研究，每一个囚犯被观察的时间不固定。要根据囚犯进入研究的时间和因释放而离开研究的时间，或者因为研究结束而无法再观察到囚犯。这个纪律犯罪率随时间变化的期间需要明确地以泊松回归模型来包含，这可以通

过包含一抵消(offset)于此模型中轻易办到。一个特定囚犯的违反率可以被模型化为 $\log_e \frac{\lambda}{t} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$，其中$\frac{\lambda}{t}$为某一特定囚犯的违反率。作为分母项的时间为观察到的纪律违反次数的时间长度。因为 $\log_e \frac{\lambda}{t} = \log_e \lambda - \log_e t$，我们可以将前面的回归方程表示为 $\log_e \lambda = \log_e t + \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$。$\log_e t$ 为抵消。对于每一个罪犯都会有一个特定值，且没有关于它的参数可以估计。如果我们将这个方程的两边指数化，就得到 $\lambda = te^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$，而且，我们可以用 t 的特定值、X_j 的特定值和未知的回归参数写出此泊松分布的对数似然。对于这个模型回归参数的解释就如同那些对每一个体有着相同固定观察时间的传统模型。

我们用 SAS 软件(SAS Institute，2002)来估计此模型。如前所述，共有11738个观察值和 14 个自变量。这个模型的对数似然为－2661.0231。此值可以用来与其他模型的对数似然相比较，在其他模型中一组自变量被舍弃以决定是否较简单的模型在预测违反率时可以像复杂模型一样好(即一个似然比检验)。此模型的离差为 25659.87。将此离差除以其自由度得到一个关于此模型拟合优度的衡量，称为尺度化的(scaled)离差。因为有 11723 个自由度(样本数 11738－估计回归中的参数数目 15)，拟合优度为 2.1888。对于泊松分布，平均数等于其方差。在这个条件下，我们会预期尺度化的离差，即我们的拟合优度测量接近于 1。如果它小于 1，则有一个过低离散(underdispersion)的情况；如果它大于 1，则有过度离散的情况。在这两种情况下，泊松分布的平均数与方差相同的条件被违反了，且模型的拟合优度也被连累(compromised)。

违反泊松分布的假设对于回归参数的估计值没有影响；然而，其对于回归系数的标准误的估计值却有影响。像我们的例子一样，过度离散(overdispersion)通常比过低离散更为常见。过度离散指的是，作为结果或因变量的计数数据——在我们的例子中为纪律违反次数——较一个泊松分布所期望的更加多变。因此，根据最大似然的标准误估计值为实际标准误的低估值，因为在似然估计方程中使用泊松平均数等于方差这一条件会低估违反次数的方差。我们

可以将它们与尺度化的离差的平方根相乘来修正原来的标准误。

同大多数回归模型一样，最重要的结果是含有估计回归参数、标准误、t 比率及显著性程度等标注的一张表格。除了截距参数估计值外，我们的泊松违反率模型包含了 14 个估计的回归参数。最重要的回归参数与结构判决(1)/公平判决(0)这一指标变量有关，因为这是研究的重点。其余的 13 个自变量为囚犯背景变量，用来作为控制变量以调整结构判决和公平判决这两个群体之间因囚犯背景差异所产生的影响效应。囚犯背景变量对于违反率的效应也应受到关注，因为它们代表了影响违反率的风险因素。

我们没有呈现所有 14 个估计的回归参数，而仅仅呈现了结构/公平判决、囚犯年龄、先前入狱及纪律违反历史这几个变量的影响效应。结构/公平判决由一个单一的指标变量所体现，1 代表结构判决，0 代表公平判决。囚犯年龄是以年数衡量的连续变量。先前入狱及纪律违反历史，如前所述，是由两个指标变量所概括的一个包括三个级别的类别变量。第一个(指标变量)将先前有服刑且违反纪律的囚犯赋值为 1，否则为 0。第二个将先前有服刑但没有违反纪律的囚犯赋值为 1，否则为 0。因此，第一个指标变量代表了先前有服刑且违反纪律的囚犯与先前没有服刑的囚犯之间的对比。第二个指标变量代表了先前有服刑但没有违反纪律的囚犯与先前没有服刑的囚犯之间的对比。因此，先前没有入狱经历的为这两个指标变量的参照组。回归分析结果呈现在表 8.1 中。

表 8.1　泊松违反(纪律)模型之参数估计值

变　量	参数估计值	标准误	t 比率	显著程度
组织对公平判决	0.2413	0.0326	7.40	<0.0001
先前有服刑且违反纪律对先前无服刑	0.5501	0.0403	13.65	<0.0001
先前有服刑但没有违反纪律对先前无服刑	0.0413	0.0341	1.21	<0.2259
囚犯年龄	−0.0831	0.0022	−37.77	<0.0001

四个变量中的三个为囚犯违反率的高度显著预测项(predictor)。结构判决囚犯与公平判决囚犯相比有显著较高的 $\log_e$(违反率)。同理，结构判决囚犯的

违反率比公平判决囚犯高。有先前服刑及违规的囚犯，相对于没有先前服刑经验的囚犯，也有着较高的违反率。囚犯年龄的估计回归参数－0.831 高度显著，表示当囚犯的年龄增大时，其违反率会降低。

估计的回归参数反映了相关自变量对违反率的自然对数的影响效应。虽然这些估计回归参数的符号和相对大小给了我们关于参数的影响效应的概念，个别估计回归参数却很难解释。因此，通常会指数化估计回归参数(即 e^{β})，以使它们可以反映在违反率上的倍数效应。计算机程序大多会输出指数化的回归参数估计值，并可以指定显著性水平(如 95%)的置信区间。

表 8.2 纪律违反率泊松回归

变　量	指数化的参数估计值	95%置信区间
组织对公平判决	1.27	[1.19, 1.36]
先前有服刑且违反纪律对先前无服刑	1.73	[1.60, 1.88]
先前有服刑但没有违反纪律对先前无服刑	1.04	[0.97, 1.11]
囚犯年龄	0.92	[0.916, 0.924]

使用表 8.1 中估计的回归参数和标准误，我们在表 8.2 中呈现了指数化回归参数及其 95%的置信区间。原本的 95%的置信区间为 $\hat{\beta}\pm1.96$ 倍 $\hat{\beta}$ 的标准误。因此，$e^{\hat{\beta}}$ 的置信区间为：

$$[e^{\hat{\beta}-1.96se_{\hat{\beta}}},\ e^{\hat{\beta}+1.96se_{\hat{\beta}}}]$$

其中 $se_{\hat{\beta}}$ 为 $\hat{\beta}$ 的标准误。

如果 $\hat{\beta}$ 的累加效应为零，则倍数效应 $e^{\hat{\beta}}=e^0=1$。因此，一个为 1 的倍数效应代表没有效应。如果倍数效应比 1 小，则相关的自变量对违反率有负的效应。如果倍数效应比 1 大，则相关的自变量对违反率有正的效应。一个负的效应是指当自变量的值上升时违反率降低。如果置信区间涵盖 1，则指数化的回归参数在给定的显著性水平上不具有统计显著性，回归参数也是如此。例如，如果一个 95%的置信区间涵盖 1，则回归参数在 0.05 水平上统计不显著。

结构/公平判决对总违反率的倍数效应为 1.27。这意味着，结构判决囚犯的违反率比公平判决囚犯的违反率高出 27%。亦即，将公平判决囚犯的违反率

乘以 1.27 就可以得到结构判决囚犯的违反率。

先前有入狱及纪律违反史的相对于没有入狱的倍数效应为 1.73。前一个群体的违反率比后者高出 73%。先前有入狱但没有违反纪律的群体与没有入狱的群体，在 0.05 的水平上，统计上没有显著差异。注意，95%的置信区间包含 1，故没有倍数效应，或者，在 $\log_e$ 尺度上没有累加效应。

年龄的倍数效应为 0.92，代表年龄每增加一年违反率会减少 8%。亦即，年龄对于整体违反率的效应是负的。注意，累加效应的参数为负（即 −0.0831）。因为年龄是连续的，我们可以检验年龄的任意增加在减少犯罪率上的效应。例如，年龄增加 10 年会有一个对于违反率 $e^{10\hat{\beta}_{age}}$ 的倍数，增加一年的效应为 $e^{\hat{\beta}_{age}}$。因为 $e^{10\hat{\beta}_{age}}$ 等于 $(e^{\hat{\beta}_{age}})^{10}$，我们只要简单地提高 $e^{\hat{\beta}_{age}}$，因增加一年而对违反率的倍数效果，到 10 次方（power）即可。我们可以使用任何计算器简单地完成这一运算。因为 $e^{\hat{\beta}_{age}}$ 为 0.92，年龄增加 10 年的倍数效应为 $(0.92)^{10}$ 或 0.43，即违反率降低 57%。

第 9 章 | 生存分析

生存分析，顾名思义，用来预测一个个人或物体生存多久，直到一个事件的发生，如在个人例子中为死亡或在物体（如一个机器的零件）例子中为失效。生存分析在所有科学研究中都有着广泛的运用（Hosmer & Lemeshow, 1999）。例如，在医学研究中，它被用来调查各种药物对癌症患者生存时间的影响。在物理科学中，它被用来模型化各种次系统（如飞机零件）的失效时间，在社会科学中也有广泛的运用。例如，它被用来模型化被解雇后找到一份新工作的时间，一个病患在退出药物滥用治疗计划前所花的时间，以及一个犯人从被监禁后发生一起监狱违规的时间。

对于这些问题以及其他类似的问题，我们想对回归模型进行改进以预测生存时间或生存时间的某种函数。也就是说，我们想要决定一组假设的自变量或协变量是否能解释生存时间或一个事件发生所需的时间。例如，知道病患和治疗计划特征与病患在自愿退出之前参与该计划的时间长度如何相关，是有用的。例如，针对参与美沙酮（methadone）治疗计划的海洛因（heroin）成瘾者的一些研究发现，接受美沙酮较高剂量的患者倾向于在治疗（计划）中持续较长的时间。当然，剂量程度以外的变量也被包含在模型中，以调整其他可能会导致治疗持续时间差异的影响，如病患的性别和年龄。

这是生存分析在前面讨论过的广义线性模型中所没有发现的一个方向，这使生存分析在某些条件下更加复杂。这个复杂的方向是删截（censoring）。在许多持续时间有限的研究中，不一定对所有的个体都有生存时间数据。在我们的美沙酮治疗例子中，可能有相当大比例的患者在研究期间不会退出治疗。这样一来，我们就不会有这些患者的生存时间，因为他们在研究结束时仍处于治疗状态。另一

个例子是一个对五年癌症患者的研究,他们到死亡发生时的生存时间为结果变量:有些患者可能在研究结束时仍然活着。虽然我们无法测量这些患者的生存时间,但我们却有这些患者的某些信息,可以在模型中用来估计回归参数。我们知道,他们存活了某段时间。我们在其后会看到如何将这个信息用于生存分析。除了直到研究结束时还存活的人,也可能有其他人失访以至于我们不知道他们的生存时间。然而,我们可以知道在某一时间点,他们还存活着(图 9.1)。

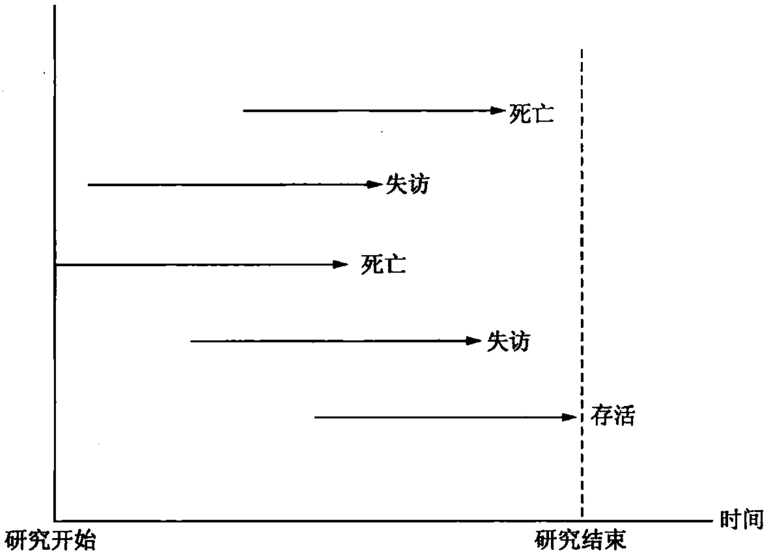


图 9.1 生存分析时间示意图

生存时间分布

在任何研究中的生存时间因个体的不同而会有所变化。它是一个连续随机变量,且如同其他任何随机变量一样有一个概率函数。许多不同的分布被用来描绘生存时间,包含威布尔、指数、gamma 以及对数正态。故需根据研究本质、理论上的正当理由及所选取的分布与实证研究数据的适切性,来选取一个特殊的分布以用于生存模型中。

许多类型的分布都允许生存时间数据有多种形状和尺度(离散)。使用最

广泛的分布为威布尔分布，因为它是一个包含了两个参数的分布(一个形状参数α和一个尺度参数λ)，根据α和λ等参数值允许多样的分布。根据参数，威布尔分布可以趋近于指数、gamma和对数正态分布的形状。图9.2画出了有着不同α值但却有相同λ值的威布尔分布。

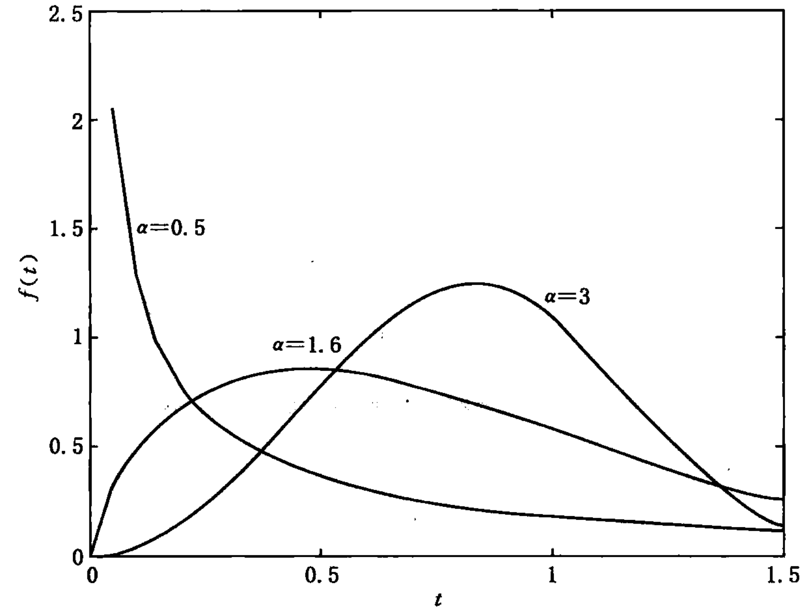


图9.2 相同尺度参数($\lambda=1.2$)但不同形状参数的三种威布尔分布

最简单的生存时间分布为单一参数的指数分布；其为$f(t)=\lambda e^{-\lambda t}$，有参数$\lambda$。因为它十分简单，所以有时会被用于生存时间的回归模型中。此回归模型的参数，如我们将见到的，易于解释，但也有一些限制，我们将在后面讨论。

现在让我们来讨论更多的关于生存时间分布的一些概念。除了通过其概率函数$f(t)$描绘出生存分布，我们也可以其分布函数$F(t)$来直接描绘它。$F(t)$为随机变量T(生存时间)的概率，等于或小于一个给定的t值。对于熟悉微积分的读者而言，此为积分：

$$F(t)=\int_0^t f(t)dt$$

$F(t)$为在密度函数$f(t)$下，由0至t左边的区域(图9.3)。

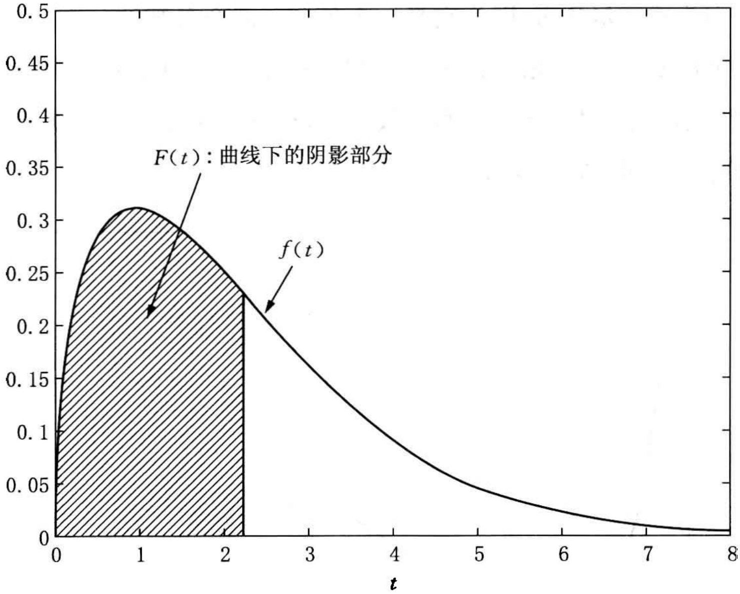


图 9.3 对于一个生存时间随机变量的密度函数($f(t)$)和分布函数($F(t)$)

一个相关的函数 $S(t)$，被称为生存函数，表示生存时间 T 等于或大于 t 的概率。因为在一个密度函数下方的区域为 1，且 $F(t)$ 为 T 小于 t 的概率，$S(t)$—

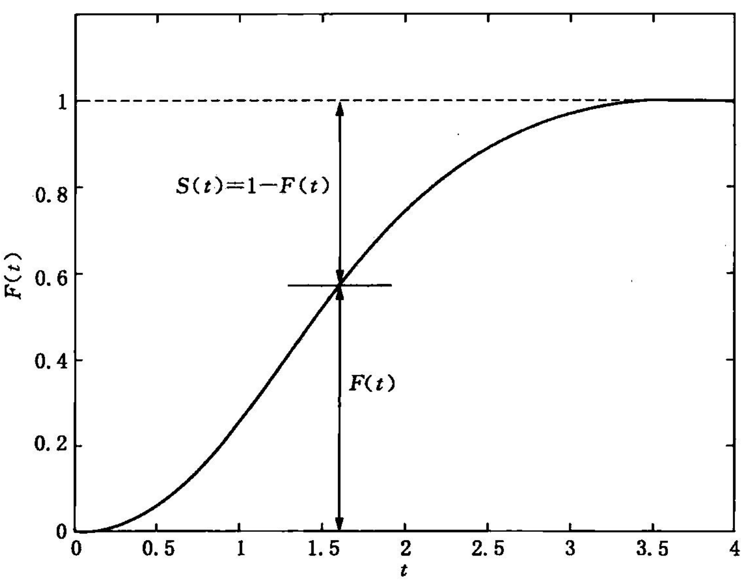


图 9.4 分布 $F(t)$ 及生存分布 $S(t)$

定等于 $1-F(t)$。这一点可见图 9.4。即，因为 $F(t)$ 为死亡或发生于时间 t 前的其他事件之概率，$S(t)$ 一定是发生在 t 和其后的事件之概率(图 9.5)。

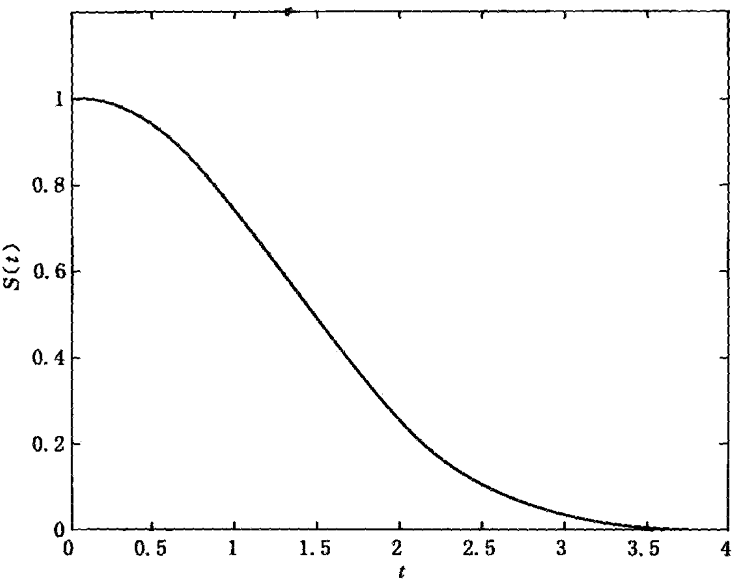


图 9.5 生存分布 $S(t)$

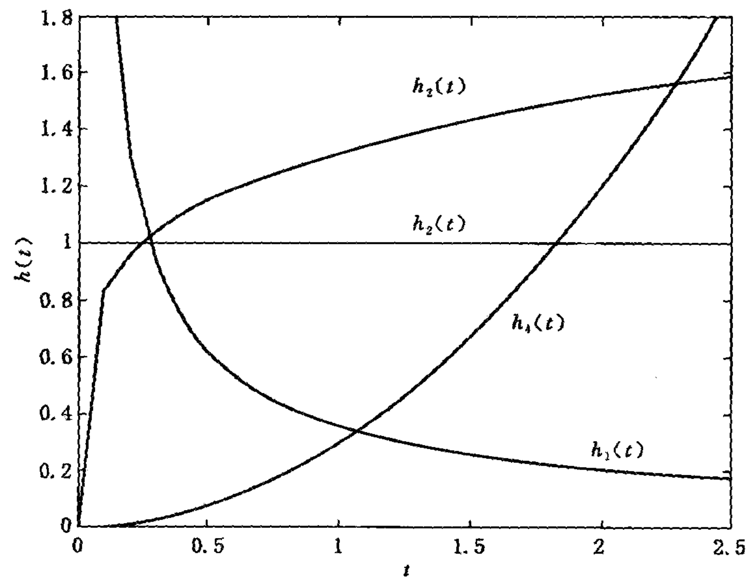


图 9.6 威布尔分布的风险函数

在生存分析中，有一个很重要的时间函数为风险函数 $h(t)$。风险函数 $h(t)$ 被定义为 $f(t)/S(t)$ 且可被解释为在给定该个体存活到时间 t 的条件下，事件发生在时间 t 的瞬间概率。它为一条件式概率且为密度函数 $f(t)$ 与生存函数 $S(t)$ 的比。如同密度函数一样，风险函数也可以有各种形式，如图 9.6 所示。

指数生存模型

最简单的参数化生存分析涉及指数分布。该密度函数，同前所述，为 $f(t) = \lambda e^{-\lambda t}$，且涉及一个单一参数 λ。相对应的生存函数为 $e^{-\lambda t}$，所以风险函数为：

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

因此，风险为一个常数，而不像在更加复杂的二参数分布（如威布尔分布）中那样（风险）是时间的函数。因为风险 λ 恒正，我们将自变量的一个线性函数模型化 $\log_e \lambda$。即，$\log_e h(t) = \log_e \lambda = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$，或 $\lambda = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$。对于未删截的数据，我们可以将这些代入对数似然函数中并求取回归参数的最大似然估计值的解。亦即，第 i 个观察值对于对数似然的贡献为 $\log_e f(t_i) = \log_e \lambda_i - \lambda_i t_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} - e^{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}} t_i$。

如果数据是删截的，则每一个观察值对于似然函数的贡献要根据第 i 个观察值所观察到的值 t_i 判断是否为删截的。对于未删截的生存时间，如前所示，对于对数似然的贡献为 $\log f(t_i)$。对于删截的生存时间，$f(t_i)$ 用于似然函数中并不适当。然而，对于删截的观察值，我们知道生存时间至少是 t_i，即使我们不知道确切的生存时间。这仍是关于回归参数的有用信息，可被用在似然函数里。因此，对于删截的观察值，可用 $S(t_i)$ 作为对似然函数的贡献。我们用一个删截指标 δ_i，对于每一个观察值，如果 $\delta_i = 1$，则 t_i 为未删截的；$\delta_i = 0$，则 t_i 为删截的，因此似然函数可以被写成：

$$\prod_i f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

或

$$\prod_i \lambda_i^{\delta_i} e^{-\lambda_i t_i}$$

其中$\prod_i$表示个别似然的乘积。对数似然为 $\sum_{i=1}^{n} \delta_i \log \lambda_i - \lambda_i t_i$。在对数似然中，将 $\log_e \lambda_i$ 替换为$\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$，并将 λ_i 替换为$e^{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}}$，并对每一个回归参数求偏导数，我们可获得一组最大似然估计方程。因为方程为非线性的，我们用迭代数值运算法(iterative numerical algorithm)来估计回归参数及其标准误。

如同所有的广义线性模型一样，我们可以从离差的差异来比较许多模型的拟合，或用似然比检定，即 $\chi^2 = -2[\ell(\hat{\beta}_1, 0) - \ell(\hat{\beta}_1, \hat{\beta}_2)]$，其中$\ell(\hat{\beta}_1, 0)$为令 β_2 等于 0 时的较简单模型的对数似然，此简单模型嵌套(nested)于对数似然为$\ell(\hat{\beta}_1, \hat{\beta}_2)$的更复杂的模型中。更复杂的模型包含额外的一组回归参数β_2。卡方的自由度等于简单与复杂模型中参数数目的差值。如果卡方为统计显著的，则简单模型被拒绝，支持更复杂的模型。如果嵌套模型只相差一个参数，我们就可以用来自完整模型的估计的回归参数及其相关标准误来执行一个 t 检验，以决定这一特定的变量是否应该被包含在模型中。

因为$\lambda_i = e^{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}}$，每一个自变量的效应为倍数的，而不像在经典多元回归中效应是累加的。对于一个正的回归参数 β_j，为相关自变量 X_j 一个单位的增加，所导致风险 λ 增加一个 e^{β_j} 的因素。如果 β_j 为负，则 X_j 每增加一个单位会导致风险降低一个 $1 - e^{\beta_j}$ 的因素。对于一个正 β_j，在 X_j 增加两个单位时会导致风险增加 $e^{2\beta_j}$ 倍等等。同样的，对于一个负 β_j，在 X_j 增加两个单位时会导致风险降低 $1 - e^{2\beta_j}$ 倍。

请注意，指数分布的平均数为 $1/\lambda$，即风险的倒数，这一点很有趣。因此，当指数分布的风险增加，平均数降低，反之亦然。我们可以用估计的回归参数来模型化第 i 个观察值的风险或指数分布的平均数，此为检验相同过程的两种方法。

指数生存模型实例

一个应用指数生存模型的经典例子是，预测一群孩子们在白血病减轻过程

中(in remission)的发作时间(Breslow, 1974)。有三个预测或预兆变量:白血球计数的自然对数(log WBC)、年龄和年龄的平方项。log WBC 被用来弥补偏态或 WBC 的界外值。使用年龄平方项是因为先前的研究显示在中间年龄范围内(即年龄与生存时间有一个曲线关系)的孩子生存时间是最长的。拟合三个模型——仅有截距(无协变量)、仅有单一变量 WBC 以及所有三个预测因素(WBC、年龄和年龄平方项)——的结果呈现在表 9.1 中。

表 9.1 指数生存模型的回归系数、对数似然值及卡方值

模型编号	预测变量	对数似然值	回归系数	χ^2	df
1	仅有截距	−1332.925			
2	log WBC	−1316.999	0.72	31.85	1
3	log WBC	−1314.065	0.67	37.72	3
	年龄		−0.14		
	年龄平方项		0.011		

资料来源:Breslow(1974)。

模型 2 的卡方借由−2(模型 1 的对数似然值 − 模型 2 的对数似然值)所获得,即 $-2[-1332.925-(-1316.999)]=31.85$:此亦等于完整和简化模型的离差的差异。卡方的自由度为 1,因为它与仅含截距的模型只差一个回归参数。故它在 0.001 水平上高度显著。当观察值数目足够大时,1 个自由度的卡方趋近于 t^2。因此,卡方为 31.85,t 值为 $\sqrt{31.85}=5.64$,且我们知道逼近 2 的 t 在 0.05 水平上是统计显著的。

模型 3 增加了年龄和年龄的平方,其卡方也高度显著($\chi^2=37.72$)。我们有理由期望这个模型显著,因为它包含 log WBC,其本身就是一个高度显著的预测项。问题是,增加年龄和年龄平方项是否相较于仅有 log WBC 时显著地增加了 FIC。我们可以将两模型对数似然的差额乘以两倍,检定关于年龄和年龄平方项的回归系数都为零的虚无假设。在卡方表中查询两个自由度所对应的卡方值为 5.87,发现在 0.05 水平上不显著。因此,我们接受虚无假设,即关于年龄和年龄平方项的参数都为零。有两个自由度是因为两个模型中参数数目的差额为 2。因此,我们接受较简单的 WBC 模型。而且,5.87 也等于两个模型

离差间的差异。

log WBC 的参数高度显著且为正(0.72)。这表示当 WBC 上升时,风险也随之上升,或者,期望生存时间随之减少。这个指数生存模型为一个比例风险模型的例子。对于指数分布固定的风险函数,会作为自变量的函数而发生变化,但会保持相同的形状——一条水平线。对于任何比例风险模型,个别协变量的重要性为 e^{β_j},而 β_j 为关于第 j 个协变量的回归参数。e^{β_j} 值代表在其他所有协变量固定的条件下,相关协变量 X_j 增加一个单位所导致的风险的倍数改变。在我们的例子中,log WBC 增加一个单位所导致的改变为 $e^{0.72}$ 或 2.05,代表一个约两倍的风险。

第10章 | 结论

广义线性模型提供弹性以使用因变量的不同概率分布来处理各种数据。它们是因变量假设为正态分布且条件平均数为自变量的线性函数的经典回归模型的普遍化。广义线性模型假设平均数的函数 $g(\mu_i)$，而非平均数，是自变量的一个线性函数。因变量可以具有指数家族中的各种分布。连结平均数函数与线性预测变量的连结函数，通常由该因变量的特殊误差分布形态决定。此被称为标准连结。对于正态分布，它是同一性(identity)连结 $g(\mu_i)=\mu_i$；对于二项分布，它是 logit 连结 $g(\mu_i)=\log_e\left(\frac{\mu_i}{1-\mu_i}\right)$；对于泊松分布，它是对数连结 $g(\mu_i)=\log_e(\mu_i)$。对于比例风险模型，风险的对数为自变量的线性函数。离差不能用来衡量一个特定模型的拟合优度。但离差间的差异可以用来比较两个备选嵌套模型的拟合。

在此书中所讨论的 GLM 假设观察值之间彼此独立。模型可以被扩展至因为观察值聚集在一个较高的分析层次内(如学校、诊所和班级)因而彼此相关的例子。这些模型被称为混合效应(或随机效应模型)，因为它们包含了牵涉回归参数的一个固定成分，再加上一个代表聚集效应的随机成分，正如我们讨论过的模型一样。该随机成分说明了聚集在同一个单位内的观察值之间的相关性。

虽然还有许多其他我们所没有讨论的广义线性模型，但与已经讨论过的模型相比，它们较少被使用。

附 录

本书所涉及的分析都是通过操作 SAS(SAS Institute, 2002)得出的。在SAS中有许多程序可以用来执行广泛的统计分析。在此,我们总结一些建立广义线性模型的重要SAS程序。

一个在SAS中非常富有弹性的程序为PROC GENMOD,它可以估计本书中讨论的所有广义线性模型。有几个内建的连结函数,包括同一性、对数、logit及probit,它也允许使用者结合七个内建的分布,即二项式、gamma、倒数、高斯、多项、负二项、正态和泊松,设定它们自己的连结函数。下面,我们提供一些如何使用此程序的例子。在这些例子中,使用者自己的说明以斜体表示,DV表示因变量,而IV表示模型中的自变量。对指令句的评论以/ * … * /表示。注意,每一个SAS指令句都需要以分号(;)作为终结。

(1) 线性回归:连结函数为同一性函数且分布为正态的。

```
proc genmod data = 数据名称;
class 如果定类自变量存在的话,其之名称;
model DV 的名称 = IV 的名称/dist = normal
                                  link = identity;
run;
```

表 3.1 报告的回归分析可以通过使用下列指令得出:

```
/ * 以下指令是为了将数据读入 SAS * /
```

```
data table31;
input y x1 x2 x3 x4 x5 x6;
datalines;
5 2 3 4 5 5 6/*这为个体1的数据,有七个字段,包含*/
/*七个在"input"指令句中定义的定序变量*/
……
;
/*我们在下面不需要"class"陈述,因为所有IV都为连续*/
/*变量*/
prog genmod data= table 31;
model y = x1 x2 x3 x4 x5 x6/dist=normal link=identity;
run;/*"run"用来执行上述SAS指令*/
```

(2) logistic 回归:连结函数为 logit 且分布为二项的。

```
proc genmod data = 数据名称;
    class 如果定类自变量存在的话,其之名称;
    model DV 的名称 = IV 的名称/dist = binomial link = logit;
run;
```

表 7.1 报告的 logistic 回归分析可以通过使用下列指令得出:

```
/*以下指令是为了将数据读入SAS*/
data table71;
    input interv JLLV JVC YM MI;
    label interv="干预(1)或无(0)"
    JLLV="青少年违法饮酒"
    JVC="暴力犯罪青少年逮捕率(每1000名青少年的被捕数目)"
```

```
    YM="年轻男性占该县人口的百分比"
    MI="该县的中位数收入(测量单位为千元)"
;
datalines;
1  250  0.64  0.25  11.5
……
;
proc genmod data= table71;
    model interv = JLLV JVC YM MI /dist = binominal link = logit;
run;
```

(3) 泊松分布:连结函数为对数且分布为泊松。

```
proc genmod data = 数据名称;
    class 如果定类自变量存在的话,其之名称;
    model DV 的名称 = IV 的名称/dist = poisson link = log
    offset=被用做抵消的变量名称;
run;
```

注意:抵消变量不可以是 DV 或 IV。

表 8.1 报告的泊松回归分析可以通过使用下列指令得出:

```
data table81;
    input infract sentence prior1 prior2 age time;
    log_time= log(time);/ * 此指令用来建立抵消变量 * /
    labe linfract = "纪律违反数"
    sentence = "组织判决(1)/公平判决(0)"
    prior1 = "先前有服刑且违反纪律(1)/先前无服刑(0)"
```

```
    prior2 = "先前有服刑但没有违反纪律(1)/先前无服刑1(0)"
    age = "囚犯年龄"
    time = "观察到纪律违反数的时间长度"
;
datalines;
4  1  1  0  24  5
……
;
proc genmod data = table81;
    class sentence prior1 prior2;
    model infract = sentence prior1 prior2 age/dist = binomial
                                              link = logit
                                              offset= log_time;
run;
```

(4) 生存分析:PROC GENMOD 不能分析删截数据或提供其他有用的生存时间分布,如威布尔或对数正态分布。然而,它可以用于对具有 gamma 分布的未删截数据建模,且可以提供指数分布相对于其他 gamma 分布选择的统计检验。

```
proc genmod data = 数据名称;
    class 如果定类自变量存在的话,其之名称;
    model DV 的名称 = IV 的名称/dist = gamma link = log;
run;
```

指数生存回归也可通过加入一个次指令 SCALE 来估计。

```
proc genmod data = 数据名称;
    class 如果定类自变量存在的话,其之名称;
```

```
    model DV 的名称 = IV 的名称/dist = gamma link = log
                                        scale = 1;
run;
```

表 9.1 报告的对数生存分析可以通过使用下列指令得出：

```
data table91;
    input onset age wbc;
    age_sq= age * age;/ * 此指令用来生成年龄的平方项 * /
    log_wbc= log10(WBC);/ * 此指令用来生成 log10(WBC) * /
    label onset = "白血病发作时间"
    wbc = "白血球计数(每千微升)"
    age = "年龄";
datalines;
8   12   14
......
;
/ * 模型 1 * /
proc genmod data = table91;
    model onset = dist = gamma
                  link = log
                  scale = 1;
run;

/ * 模型 2 * /
proc genmod data = table91;
```

```
    model onset = log_wbc / dist = gamma
                            link = log
                            scale = 1;
run;

/* 模型 3 */
proc genmod data = table91;
    model onset = log_wbc age age_sq / dist = gamma
                                       link = log
                                       scale = 1;
run;
```

为了获得在第 6 章后面叙述过的预测值、pearson 和离差残差，我们可以用次指令来生成这些值，以评估模型的拟合优度。以泊松回归为例进行说明：

```
proc genmod data = 数据名称;
    class 如果定类自变量存在的话，其之名称;
    model DV 的名称=IV 的名称/dist = poisson link = log;
    offset=被用做抵消的变量名称
    output out=使用者指明之输出文件名称
    pred=使用者指明之预测值名称
    reschi=使用者指明之生成 Pearson 残差名称
    resdev=使用者指明之生成 deviance 残差名称;
```

为了打印出预测值和残差，

```
proc print data = 由上述生成次指令得到的生成变量名称;
```

PROC GENMOD 也可以被用来分析历时的或聚集的数据，因为篇幅限制，

我们在本书中没有讨论。也有其他的 SAS 程序可用来估计广义线性模型的特殊种类。我们仅列出一些一般的用法供读者参考。

对于线性回归

PROC REG 和 **PROC GLM** * :这两个程序都可以用来拟合线性回归并允许定类和连续自变量。对 **REG** 程序,使用者必须建立虚拟(dummy)变量以代表定类自变量。对于 GLM 程序,使用者不用这么做,但必须要于"**CLASS**"指令句,即 **GLM** 的一个次指令,说明定类自变量。

* **GLM** 这里代表所有的广义线性模型。

对于 logistic 回归

PROC LOGISTIC:可以拟合二项的或定序的(ordinal)结果之 logistic 回归。也可以提供许多模型建构方法并计算许多回归参数诊断。

对于 Probit 回归

PROC PROBIT:可以执行 logistic 回归、定序的 logistic 回归及 probit 回归。当因变量是二分的(dichotomous)或多分的(polychotomous)而自变量为连续的,**PROBIT** 程序很有用。probit 回归与 logistic 回归相似,除了其连结函数为正态(高斯)而非 logit。

对于生存回归

PROC PHREG:可以执行基于 Cox 比例风险模型的生存数据之回归分析,其假设了对于解释变量效应的参数形式,但没有指明生存函数的本质形式。它也允许不能于 **PROC GENMOD** 中处理的删截生存时间观察值。如果有区间删

截的(interval-censored)观察值(确切的生存时间没被观察到,仅知道某一个区间),则可以使用 **PROC LIFTEST** 代替。

有一个新开发的程序 **PROC QLIM** 可以估计单变量和多变量 logit 和 probit 模型。

参考文献

Breslow, N. (1974). Covariance analysis of survival data under the proportional hazards model. *International Statistics Review*, *43*, 43-54.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: John Wiley.

Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modeling based on generalized linear models*. Berlin: Springer-Verlag.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis*. New York: John Wiley.

Le, C. T. (1998). *Applied categorical data analysis*. New York: John Wiley.

Lee, E. T. (1992). *Statistical methods for survival data analysis* (2nd ed.). New York: John Wiley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: John Wiley.

Nelder, J. A., & Wedderburn, T. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society (Series A)*, *135*, 370-384.

SAS Institute, Inc. (2002). *SAS/STAT 9 user's guide* (Vols. 1-3). Cary, NC: Author.

译名对照表

alternative	备择
binomial	二项的
binary	二元
censor	删截
component	成分
canonical	标准
conditional mean	条件平均数
covariate	协变量
data	数据
dependent variable	因变量
dispersion	离散
distribution	分布
effect	效应
error	误差
estimate	估计值
estimator	估计量
factor	因素
fit	拟合
gaussian	高斯
generalized linear model	广义线性模型
goodness of fit	拟合合优度
heteroscedasticity	异方差性
identity	同一性
independent variable	自变量
lack of fit	失拟
least square	最小二乘
link	连结
likelihood	似然
normal distribution	正态分布
poisson	泊松
predictor	预测(量)
procedure	程序
probability	概率
proportional	比例
ratio	比、比率
residual	残差
skewed	偏斜的
standard deviation	标准差
standard error	标准误
variance	方差
Weibull	威布尔

应用 logistic 回归分析

作 者 简 介

斯科特·梅纳德(Scott Menard)

美国科罗拉多大学行为科学院副研究员。毕业于康奈尔大学社会学系,并于科罗拉多大学获社会学博士学位。主要研究兴趣是:滥用药物和其他违法行为的追踪式研究。出版书籍包括:Sage QASS 系列的《追踪研究》(1991; 2002)、《透视出版》(*Perspectives on Publication*)(与伊莉沙白·莫恩合著,1987)、《问题青年》(*Multiple Problem Youth*)(与埃利奥特等人合著,1989)、《童党》(*Juvenile Gangs*)(与科维等人合著,1997)。

译 者 简 介

李俊秀

本科毕业于香港中文大学统计系,2006 年获美国纽约州立大学奥本尼分校教育心理学与方法论硕士学位。现任香港科技大学应用社会经济研究中心项目经理(project officer)。研究兴趣包括:教育、健康和定量研究方法。

序

线性回归模型是一个非常有效且重要的数据分析方法。研究人员主要着重解释因变量，将因变量看做是由多个自变量 X_1 至 X_k 所组成的函数。当所有线性回归假设都符合该模型时，模型辨识、变量测量、普通最小二乘法(Ordinary Least Squares, OLS)估计方程，这一切都很顺利。可是，当因变量有两个或三个分类的话，有几项假设就不符合了。以二分因变量为例，同方差、线性和正态性的假设都不能成立，OLS 的估计也无效。logistic 回归的最大似然估计法(maximum likelihood estimation) 就能解决这一问题，即将 $Y(1, 0)$转化成 logit(发生比的对数 log)。

梅纳德教授全面地解释了 logistic 回归模型的估计、解释和诊断结果。为了令读者能够从熟悉的事件过渡到新事物，他系统地把 logistic 回归与线性回归模型的 OLS 的 R^2、估计标准误差(standard error of estimate)、t 比率(t ratio)和斜率(slope)做比较。传统回归诊断和学生化残差(studentized residual)、杠杆(leverage)、dbeta 都包括在创新的 logistic 诊断法内。最后仔细说明了多选项和不排序多分类因变量的问题。

本书讨论了对最新计算机软件的应用，如 SPSS10 NOMREG 用以分析类别变量(nominal variable)比较好，SAS lOGISTIC 分析定序因变量比较好。本书更新了现今应用的计算机软件，同时深入地评论了不同的拟合优度(goodness of fit)。梅纳德博士还提出令人信服的论据去说明 R_L^2 的优势，至少这能直接与 OLS 的 R^2 比较。他同时增加了新内容：分组数据(grouped data)、预测效率(predictive efficiency)和风险比(risk ratios)。

大量著作证明了线性回归的广泛应用，可是由于现实中的因变量很少会是连续的(continuous)或定距的(interval)，因此，logistic 回归开始备受关注。首先出版的是德马里斯(DeMaris)的《对数模型》；接着是梅纳德的《应用 logistic 回归》(第一版)；以及潘帕(Pamper)的《logistic 回归简介》)本书从基本原理到技术应用都做了介绍，除此之外，还提及了当今最复杂的问题和方法。社会科学家要熟悉日新月异的知识，千万别错过梅纳德的这本书。

——迈克尔·刘易斯—贝克(Michael Lewis-Beck)

第 1 章 | 线性回归和应用 logistic 回归模型

只要两个变量的关系能用方程式 $Y=\alpha+\beta X$ 来表达，那就有可能用线性回归分析去检视两者是否有线性相关，同时计算此相关的强度。Y 是被预测的变量，称为因变量（dependent variable）、基准变量（criterion variable）、结果变量（outcome variable）或内生变量（endogenous variable）；X 是用来预测 Y 的变量，称为自变量（independent variable）、外生变量（exogenous variable）或预测变量（predictor variable）[1]；α 和 β 是总体参数（population parameters）的估计值，参数 α 称为截距（intercept），代表 $X=0$ 时 Y 的数值；参数 β 代表 X 增加一个单位时 Y 数值的变化，或是 X 预测 Y 的最佳直线斜率。多元回归（multiple regression）含有几个自变量，假设 K 是自变量的数量，方程式便是 $Y=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k$，而 β_1，β_2，…，β_k 称为偏斜率系数（partial slope coefficient），即任何一个自变量 X_1，X_2，…，X_K 只对 Y 值提供部分预测。有时方程式会明确地显示出 X 对 Y 的预测并不精确，$Y=\alpha+\beta X+\varepsilon$，数个自变量的方程式 $Y=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon$，$\varepsilon$ 代表 X 预测 Y 时出现的误差，是一个随机变量。就个别个案 j，$Y_j=\alpha_j+\beta_jX_j+\varepsilon_j$ 或 $Y_j=\alpha_j+\beta_1X_{1j}+\beta_2X_{2j}+\cdots+\beta_kX_{kj}+\varepsilon_j$，$j$ 表示指定的第 j 个个案（$j=1$ 代表第一个个案，$j=2$ 是第二个个案，如此类推）。Y_j、X_{1j}、X_{kj} 等代表因变量和自变量的特定值。上述的方程式主要用来计算在个案 j 中，Y 的数值，多用于描述变量间的关系。

截距 α 和回归系数 β（regression coefficient）的估计值是通过普通最小二乘法（OLS）计算出来的，这点在许多基础统计学的书中都讨论过（Agresti & Finlay，1997；Bohrnstedt & Knoke，1994）。这些估计值所形成的方程式 $\hat{Y}=a+bX$（一个自变量）或 $\hat{Y}=a+b_1X_1+b_2X_2+\cdots+b_kX_k$（多个自变量）中，

$\hat{Y}$ 是 Y 的线性回归方程式的预测值，a 是截距 α 的普通最小二乘法的估计值，b（或 b_1，b_2，…，b_k）是斜率 β（或 β_1，β_2，…，β_k）的普通最小二乘法的估计值。每个个案的残差(residuals) $e_j = (Y_j - \hat{Y}_j)$，其中 $\hat{Y}_j$ 是 Y 在个案 j 的估计值。二元回归(bivariate regression)的残差可以用二元散点图表示，即每一点和回归直线的垂直距离。多元回归的残差就比较难，因为它需要多维空间才能展示。

图 1.1 是二元回归模型的一个例子，其数据来自 1980 年第五次美国家庭调查，访问对象是 16 岁的青少年。图 1.1A，因变量 FROMRJ5，即被访者自我报告的每年使用大麻的频率（在过去一年，你曾经吸过多少次大麻?）；自变量 EDF5，即接触违法朋友的程度[2]。量度接触违法朋友的方法是将八个问题加起来，问受调查的青少年到底有多少个朋友曾被牵涉在不同的罪行当中（如偷窃、侵犯他人、药物滥用）？每题有五个选项，由 1(没有朋友)至 5(所有朋友)，因而 EDF5 的总和是由 8 至 40。图 1.1A 显示接触违法朋友与大麻使用成正相关，方程式如下：

$$(\text{FROMRJ5}) = -49.2 + 6.2(\text{EDF5})$$

换言之，接触违法朋友每增加一个单位，每年大麻使用频率就会随之增加约六倍，或每两个月增加一次。

决定系数(coefficient of determinant)R^2 是指自变量能多准确地预测因变量。通过知道青少年接触违法朋友的数量，我们可根据 EDF5 的值和回归方程去表现它与 FRQMRJ5 之间的关系。这一方法可减少 12%的预测平方误差总和(sum of the squared errors of prediction)

$$\sum e_j^2 = \sum (\hat{Y}_j - Y_j)^2,\ R^2 = 0.116$$

当解释结果时，必须考虑因变量和自变量的真实数值。截距指的是没有接触过违法朋友的人，其大麻使用频率为负数。出现这种不合理的数值是因为接触违法朋友的数值由 8(完全没有朋友涉及任何一项违法活动)至 40(所有朋友都涉及违法活动)。因此，当个别人的接触违法朋友的数值最小时，大麻使用的期望值应为 $-49.2 + 6.2(8) = 0.4$，接近于零，表示即使没有接触违法朋友的

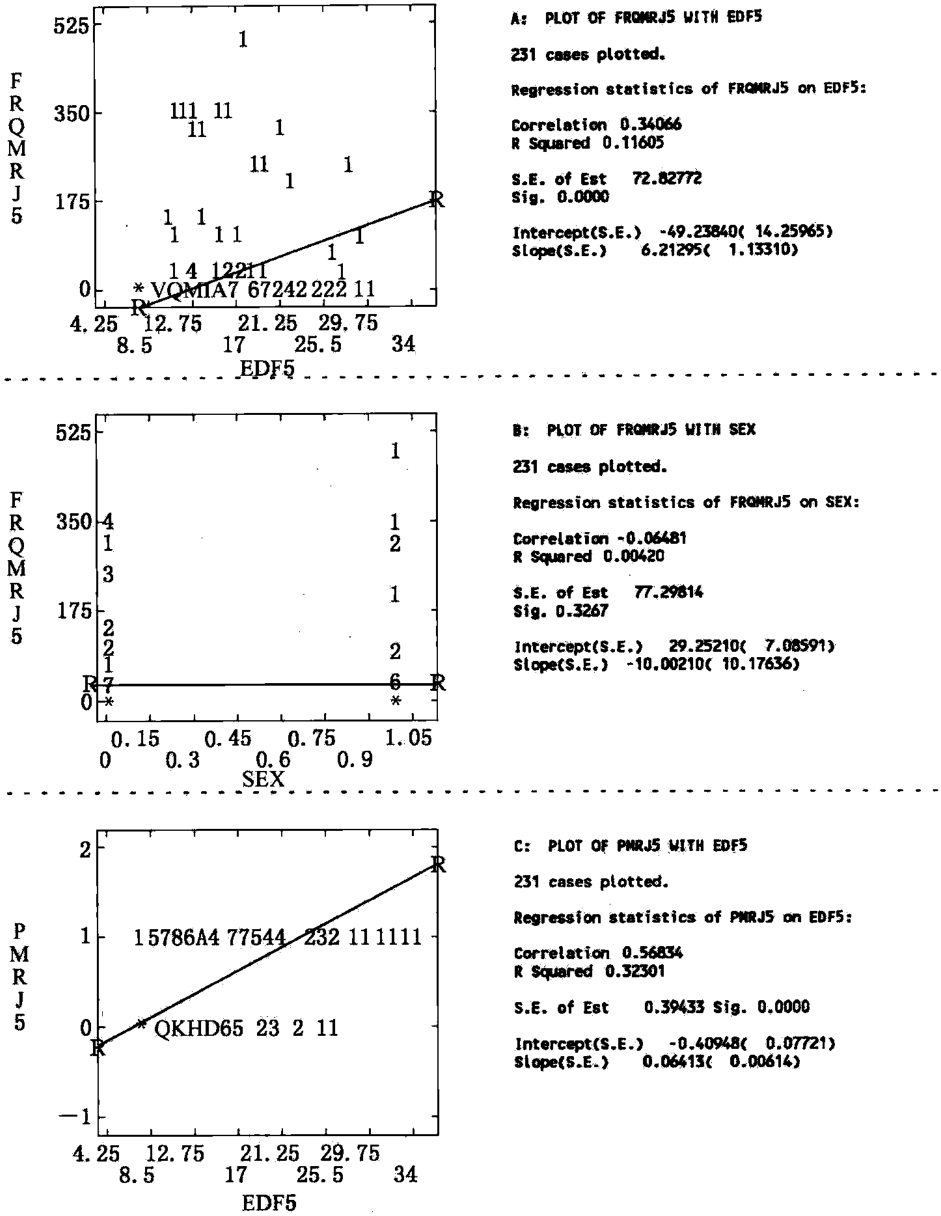


图 1.1 二元回归图

人，偶尔也会吸食大麻。这一研究样本的 EDF5 最大值是 29，相应的大麻使用的期望值为 $-49.2+6.2(29)=130.6$ 或每三天吸食一次大麻。这样，无论是从统计学或现实世界的角度看，这一数值都算合理。

回归假设

运用 OLS 估计或推算线性回归分析中的系数，必须符合以下假设：

1. 测量(measurement)：所有自变量是定距(interval)、比例(ratio)或二分的(dichotomous)，因变量是连续的(continuous)，没有限度，以及定距或比例尺度。所有变量没有量度误差[3]。

2. 模型设定(specification)：(1)分析中要包括与因变量相关的所有自变量；(2)分析要包括因变量的所有不相关自变量；(3)线性相关(自变量和因变量可做转换)。

3. 误差的期望值：误差的期望值 $\varepsilon = 0$。

4. 同方差(homoscedasticity)：自变量的所有数值的误差、方差都相同或是常数。

5. 误差正态性(normality)：自变量的每组数值的误差呈正态分布。

6. 非自相关：自变量的不同值所产生的误差没有自相关。数学式为 $E(\varepsilon_i, \varepsilon_j) = 0$。

7. 自变量与误差项不相关：自变量与误差没有相关性。数学式为 $E(\varepsilon_j, X_j) = 0$。

8. 不存在多元共线性问题：在多元回归中，自变量之间不能出现完全的相关性。从数学的角度来说，对于任何一个 i，当 $R_i^2 < 1$ 时，R_i^2 是自变量 X_i 与其他自变量(X_1, X_2, …, X_{i-1}, X_{i+1}, …, X_k)的方差。当然，如果只有一个自变量，多元共线性就不成问题。

违反测量假设(measurement assumption)：二分变量的线性回归

线性回归模型可以轻易地延伸到二分预测变量，包括虚拟变量(dummy variables)(Lewis-Beck，1980：66—71；Berry & Feldman，1985：64—75；Hardy，1993)。在图 1.1B 中，因变量仍旧是自我报告的每年大麻使用频率，但这次的自变量是性别(编号 0=女性，1=男性)。回归方程式是：

$$(FROMRJ5) = 29.3 - 10.0(SEX)$$

此图包括两栏：一栏代表女性的大麻使用频率，另一栏代表男性的大麻使用频率。二分自变量的编号是 0 或 1，因此截距和斜率具有特殊的意义。

截距是当自变量为零时，因变量的预测数值（大致上是女性被访者），但只有两组的话，截距是组别编号为零（女性）的个案的平均大麻使用率。斜率仍然是当自变量增加一个单位 Y 值的变化，但只有两类的话，该数值就变成第一组（女性）和第二组（男性）的平均差。斜率和截距的总和，$29.3 - 10.0 = 19.3$，就是第二组（男性）的平均大麻使用率。图 1.1B 显示：女性的大麻使用率比男性高，但差距在统计上不显著（显著性是 0.326）。在图 1.1B 中，回归线只是连接男性与女性的平均大麻使用率的一条线，分别是女性与男性大麻使用的条件平均数（conditional mean）[4]。

当因变量是二分的，回归方程的解释就不那么直接。在图 1.1C 中，自变量是接触违法朋友，因变量是大麻使用率：在过去一年是否吸食过大麻（是＝1，否＝0）。图 1.1C 出现两条平行的水平线，不像图 1.1B 的两栏数值。二元因变量（编号 0 或 1）的线性回归模型，称为线性概率模型（linear probability model）（Agresti，1990：84；Aldrich & Nelson，1984）。图 1.1C 的方程式为：

$$PMRJ5 = -0.41 + 0.064(EDF5)$$

当因变量是二分时，其平均数就是某个个案落在两个分类中较高数值的概率函数[5]。将变量的分类编号 0 和 1，这意味着变量的平均值就是落在较高数值的个案比率。还有，因变量的预测值（在 X 的特定值和假设 X 和 Y 是线性相关的条件平均值）可解释为预测概率（predicted probability），也就是指某个个案在特定的自变量值下，落在因变量较高分类数值的概率。在理想状态下，我们希望预测概率在 0 至 1 之间，因为概率不会小于 0 或大于 1。

据图 1.1C，因变量的预测值可能高于或低于其本身的可能值。若 EDF5 为最小值（$EDF5 = 8$），预测大麻使用率（即大麻使用预测概率）便是 $-0.41 + 0.06(8) = 0.10$，这是一个合理的结果。但若 EDF5 为最大值（$EDF5 = 29$），预测大麻使用概率便是 $-0.41 + 0.064(29) = 1.45$，不可能存在一个接近 1.5 的

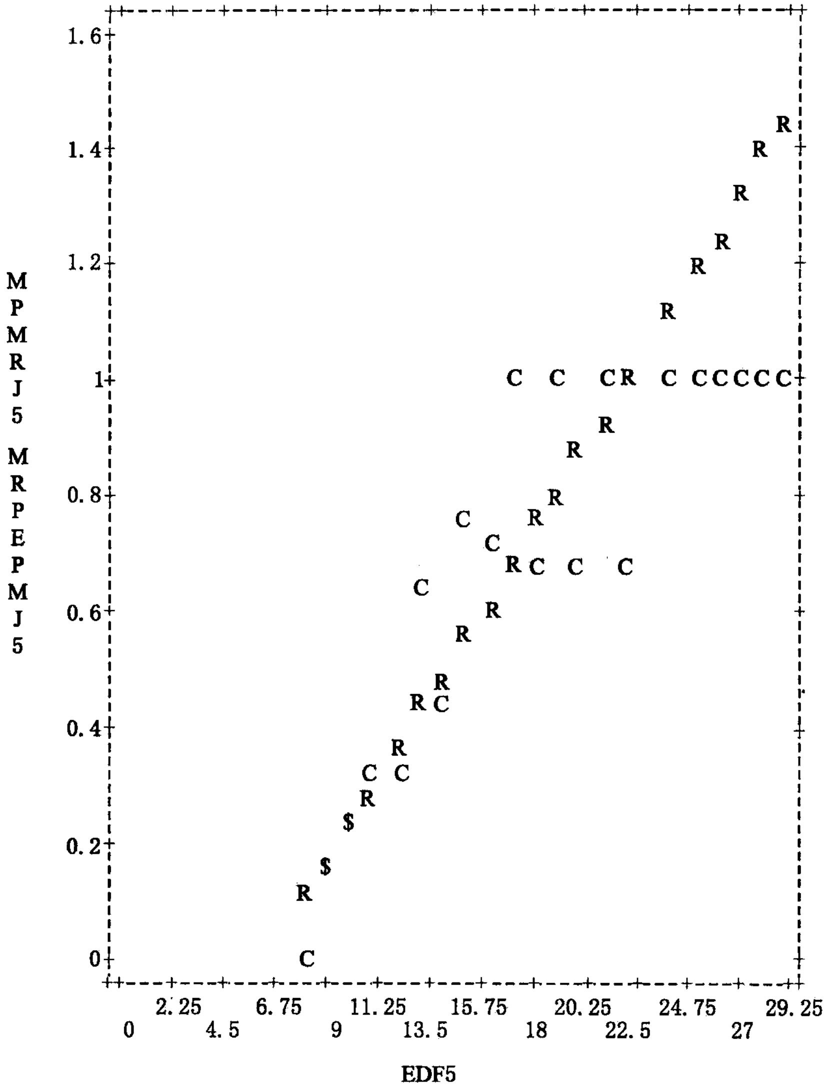
概率。而且残差的变异性随着自变量的数目而变(Schroeder, Sjoquist & Stephan, 1986:79—80; Aldrich & Nelson, 1984:13),这种情况称为异方差(heteroscedasticity),它表示虽然回归系数估计并没有偏差(没有系统性的过高或过低),但是从标准误差最小化的角度来看,这不是最好的估计。根据不同的 X 值,残差值会出现系统性的模式。参照图 1.1C,当 X 值大于 23.5,所有残差值便是负数,因为 $\hat{Y}_j$ 大于 Y_j(因为当 X 大于 23.5, $\hat{Y}_j$ 大于 1,但 Y_j 则小于或等于 1)。另外,由于残差不呈现正态分布(Schroeder et al., 1986:80),因而样本方差的估计也不会准确(Aldrich & Nelson, 1984:13—14),因此,回归系数的假设检验(hypothesis testing)或置信区间(confidence interval)都不会有效。

非线性(nonlinearity)、条件平均数(conditional means)和条件概率(conditional probabilities)

当 X 与 Y 是线性相关且有某个特定的 X 值时,连续因变量 Y 的回归估计 $\hat{Y}$ 可视为 Y 的条件平均值的估计。在二元回归中,若自变量是连续的,Y 的估计值未必完全等于 Y 的平均值,因为对于不同的 X 值,Y 的条件平均数未必准确地落在一条直线上。若自变量是二分变量的话,回归线会穿过 X 的每个分类的对应 Y 的条件平均数。如果把 FROMRJ5 的条件平均数对二分预测变量 SEX 画出来,该图上会有两点:男性和女性的 Y 的条件平均数(注意,个案是根据自变量的数值聚集在一起的)。最简便的方法就是把这两点用直线连在一起,同时会得到不错的线性回归模型。

图 1.2 显示二分因变量的内在非线性关系。图 1.2 是 PMRJ5 的观测条件平均数(即是大麻使用率)与自变量 EDF5 间的关系,字母"C"代表观测条件平均数。因为 PMRJ5 的编号为 0 或 1,所以条件平均数代表 0 和 1 的平均数,这可被理解为条件概率。因此,图 1.2 是 PMRJ5 = 1 对不同的 EDF5 值的概率图。在图 1.2 中,所有 Y 的观测值都在 0 和 1 两条垂直线之间。从原则上讲,如果我们利用线性概率模型,预测概率就可以是无限大或无限小。

据图 1.1C 的回归方程(R),图 1.2 的观测条件概率(C)图叠加于预测条件



注:C:大麻使用率的观测平均值(MPMRJ5)对 EDF5 接触违法朋友。
R:线性回归预测的大麻使用率(MRPEPMJ5)对 EDF5 接触违法朋友。
$:重复出现(线性回归预测和观测一致)。21 个个案。

图 1.2 观测条件概率(C)和线性回归预测的条件概率(R)

概率图之内。当 EDF5 大于 23.5 时,由于条件大麻使用率的平均观测值会在 PMRJ5 = 1 停止,而回归方程的预测值于 PMRJ5 = 1 后仍继续增加至最大值 1.45,因此,当 EDF5 由 23.5 增至最大值 29 时,其预测误差一直在增加。

图 1.2 中有两点值得注意。第一，虽然线性模型可能适用于连续性的因变量，不管自变量是连续的还是二分的，证据都显示非线性回归更适合二分变量 PMRJ5。一般而言，对很高的 X 值（或很低时，如果为负相关）来说，条件概率 $Y=1$ 将会非常接近 1，而且当 X 继续增加时，条件概率的改变会较小。图 1.2 解释了这点。同时，X 值很低（或很高时，如果为负相关），条件概率 $Y=1$ 将会非常接近 0，而且当 X 继续减少时，条件概率的改变会非常小。因此，代表 X 和 Y 的关系的曲线应变得平滑，当 X 值很高和很低、斜率接近于 0 时，原则上是趋向无限大或无限小。如果 X 和 Y 相关，在很高与很低的 X 值的中间，曲线斜率就会变得更陡些，与 0 有明显的区别，通常会出现图 1.3 的 S 曲线图。

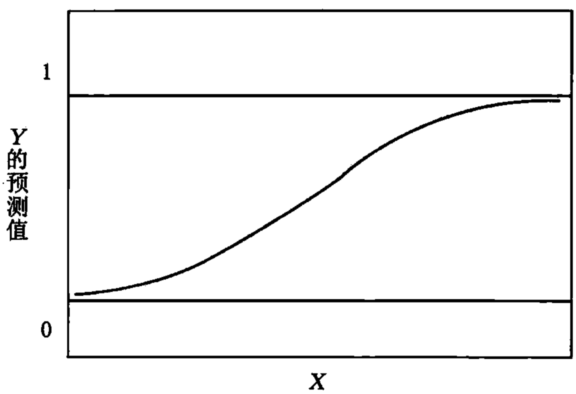


图 1.3　二分因变量的 logistic 曲线模型

第二，对整体数据而言，Y 的观测条件平均值等于 $Y=1$ 的观测条件概率，同时，Y 的预测值等于 $Y=1$ 的预测条件概率。Y 的两个类别的数值是随意编写的，可以是 0 和 1 或 2 和 3（在这种情况下，Y 的预测值等于 2 加上条件概率 $Y=3$，仍然是条件概率函数，而在特定 X 值下，Y 值为其两个数值中较高的一个）。Y 本身的数值并不重要，然而，Y 值为其两个可能值的其中一个或另一个的概率在一定程度上依赖于一个或多个自变量。

利用 OLS 线性回归去分辨 Y 随意编号数字与 Y 出现其两个值其中一个的概率是很困难的。因此，我们需要思考用别的方法去描述 X 和 Y 的关系并估计其参数。首先是非线性。当连续自变量 X 和因变量 Y 出现非线性关系时，可以

应用非线性转换的方法(Berry & Feldman, 1985)。相似的方法也可应用于二分因变量。

非线性关系和变量转换

当变量的关系呈非线性时,可转换因变量或一个或多个自变量,这时因变量与自变量的实质关系仍是非线性的,只是形式上呈线性,可以用 OLS 估计来分析。换句话说,虽然变量间的关系是非线性的,但参数间的关系可以是线性的(Berry & Feldman, 1985:53)。贝里和费尔德曼(Berry & Feldman, 1985:55—72)和刘易斯—贝克(Lewis-Beck, 1980:43—47)都曾举例说明通过变量转换而形成的线性关系。

图 1.2 有很多证据显示了大麻使用频率和接触违法朋友呈非线性相关。其中一个可能的转换是将因变量 FRQMRJ5 做对数变换[6],即加 1 后取自然对数(加 1 是为了避免零的自然对数)。回归方程为 $\ln(Y+1)=\alpha+\beta X$, 或 $(Y+1)=e^{\alpha+\beta X}$ 或 $Y=e^{\alpha+\beta X}-1$, $e=2.72$ 是自然对数的基数。在上述例子中,大麻使用频率和接触违法朋友的关系表达式为:

$$\ln(\text{FRQMRJ5}+1)=-1.7+0.23(\text{EDF5}),\ R^2=0.32.$$

比较图 1.1A 未转换的模型和对数变换模型的结果,其斜率仍为正数,但数值不同(因为因变量的单位由频率改为对数频率)。

当因变量转换后,转换模型的决定系数由 0.12 增加至 0.32,反映出这种线性模型较好。证据(不是结论,只是证据而已)显示大麻使用频率和接触违法朋友是非线性的关系。相同的结果也出现在二分预测变量(dichotomous predictors)/自变量—性别和大麻使用频率的关系中。在转换方程式中,因变量做对数变换,可解释方差由微弱的 0.004 增加至普通的 0.028,性别和大麻使用频率的关系也在统计上呈显著 ($p=0.011$)。虽然大麻使用频率和两个预测变量都是非线性关系,但是,我们仍可用线性模型和 OLS 去估计模型中的参数。

二分因变量的概率、发生比、优比和 logit 转换

正如上文所述，二分因变量的数值是随意加上的数字，数字本身并不是问题的真正趣味所在。真正有趣的是个案的分类，即个案是属于因变量的哪个组别？是否能被自变量所预测？与其预测分类组别的随意数值，更有用的方法是重构问题，预测个案被分类到两组之中的一组（与另一组相反）的概率。因为分到第一或较小数值的组别（$P[Y=0]$）的概率等于 1 减去分到第二或较大数值的组别（$P[Y=1]$）的概率，所以如果我们知道其中一个概率，就会知道另一个：$P(Y=0)=1-P(Y=1)$。

虽然我们可利用 $P(Y=1)=\alpha+\beta X$ 得出 $Y=1$ 的概率，但随之又出现另一个问题，$P(Y=1)$ 的观测值必须在 0 至 1 之间，预测值却可能小于 0 或大于 1。解决方法是把 $Y=1$ 的概率换成 $Y=1$ 的发生比。$Y=1$ 的发生比，写成 odds$(Y=1)$，即是 $Y=1$ 的概率与 $Y\neq 1$ 的概率之比。$Y=1$ 的发生比等于 $P(Y=1)/[1-P(Y=1)]$。与概率 $P(Y=1)$ 不同，发生比没有固定的最大值，但有最小值 0。

原则上，进一步转换发生比的变量，其范围是由负无限至正无限。发生比的自然对数（natural logarithm），$\ln\{P(Y=1)/[1-P(Y=1)]\}$，称为 Y 的 logit，写成 logit(Y)，当发生比由 1 减至 0，在 logit(Y)变成负数的同时，其绝对值会增加；当发生比由 1 增至无限，logit(Y)向正数方向增加。如果我们采用 $Y=1$ 的发生比的自然对数作为因变量，就不会遇到估计概率超出概率的最高或最低范围的问题。因变量和自变量的方程式为：

$$\text{logit}(Y)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \qquad [1.1]$$

我们可以以指数方式（exponentiation）将 logit(Y)转换成发生比，计算出 $\text{odds}(Y=1)=e^{\text{logit}(Y)}$。其结果是：

$$\text{odds}(Y=1)=e^{\ln[\text{odds}(Y=1)]}=e^{(\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)} \qquad [1.2]$$

$Y=1$ 的发生比是随着 X 每一单位的改变乘以 e^{β}。然后，我们可应用算式

P(Y = 1) = (Y = 1 的发生比)/[1+(Y = 1 的发生比)]将发生比转换成概率，即 Y = 1 的概率等于 Y = 1 的发生比除以 1 加上 Y = 1 的发生比。方程式为：

$$P(Y=1)=\frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}} \qquad [1.3]$$

概率、发生比和 logit 是以三种不同方式去表达同一样东西。在这三种量度中，概率或发生比可能最易理解。在数学中，概率的 logit 能最有效地帮助我们分析二分因变量。正如我们对连续因变量(大麻使用频率)做自然对数去修正大麻使用频率与接触违法朋友的非线性关系一样，我们也可以对二分因变量(大麻使用频率)做 logit 转换，去修正大麻使用频率与接触违法朋友的非线性关系。

任何一个特定的个案，logit(Y) =±∞。这保证了模型(方程式 1.3)的估计概率不会小于 0 或大于 1，但它也意味着因为线性模型(方程式 1.1)中的因变量有无限大或无限小的数值，OLS 无法预测参数。取而代之，最大似然法(maximum likelihood)能充分利用函数值，在指定自变量和参数 α，β_1，β_2，…，β_k 中，对数似然(log-likelihood)函数可显示得到 Y 观测值的可能性。与 OLS 可直接算出参数的答案不同，logistic 回归模型先用一个暂定的答案，稍做修订后再重新估计，看看是否有改善，不断重复以上过程直至其中一个步骤与下一步骤改变的结果差异微小到可被忽略为止。这个估计、检验、再估计的重复过程称为迭代(iteration)，从重复估计而得出一个答案的过程称为迭代过程(iterative process)。当似然函数从一步到下一步所做的改变可被忽略，即称为收敛(converge)。所有过程由特定设计的计算机程序处理，去查找指定的最好参数集以最大化对数函数。当 OLS 的假设完全符合时，OLS 线性估计的参数与最大似然法的估计是一致的(Eliason，1993:13—18)。换言之，OLS 是最大似然法的一个特定的情况，也可以无须迭代而直接得出一个答案的估计方法。

logistic 回归：导论

图 1.1C 显示大麻使用频率(PMRJ5)与接触违法朋友(EDF5)的 OLS 线性回归的分析结果。图 1.4 展示这两个相同变量的二元 logistic 回归的分析结果。

这是 SPSSlogistic 回归的输出结果：

```
logistic regression pmrj5 with edf5/method=enter edf5/method=bstep(lr)/print=def/classplot/
      save=pred(lpepmrj5).
```

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	85.359	1	0.000
	Block	85.359	1	0.000
	Model	85.359	1	0.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	213.947	0.309	0.425

Classification Table(a)

	Observed		Predicted		
			PMRJ5		Percentage Correct
			no	yes	
Step 1	PMRJ5	no	136	14	90.7
		yes	37	44	54.3
	Overall Percentage				77.9

a The cut value is 0.500

Variables in the Equation

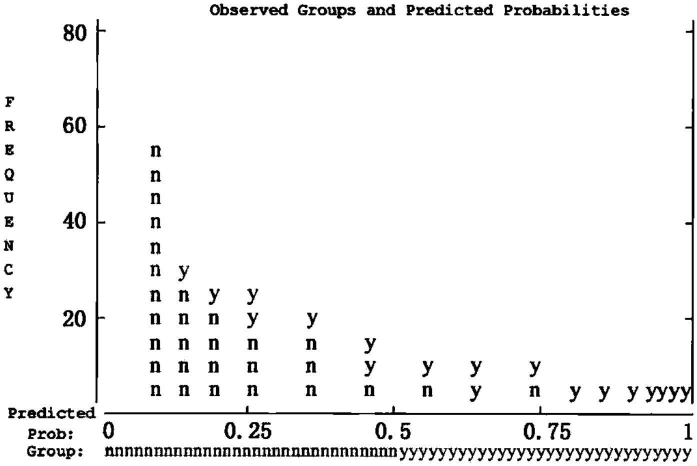
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDF5	0.407	0.058	48.546	1	0.000	1.502
	Constant	-5.487	0.710	59.732	1	0.000	0.004

a Variable(s) entered on step 1: EDF5.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.875	6	0.130

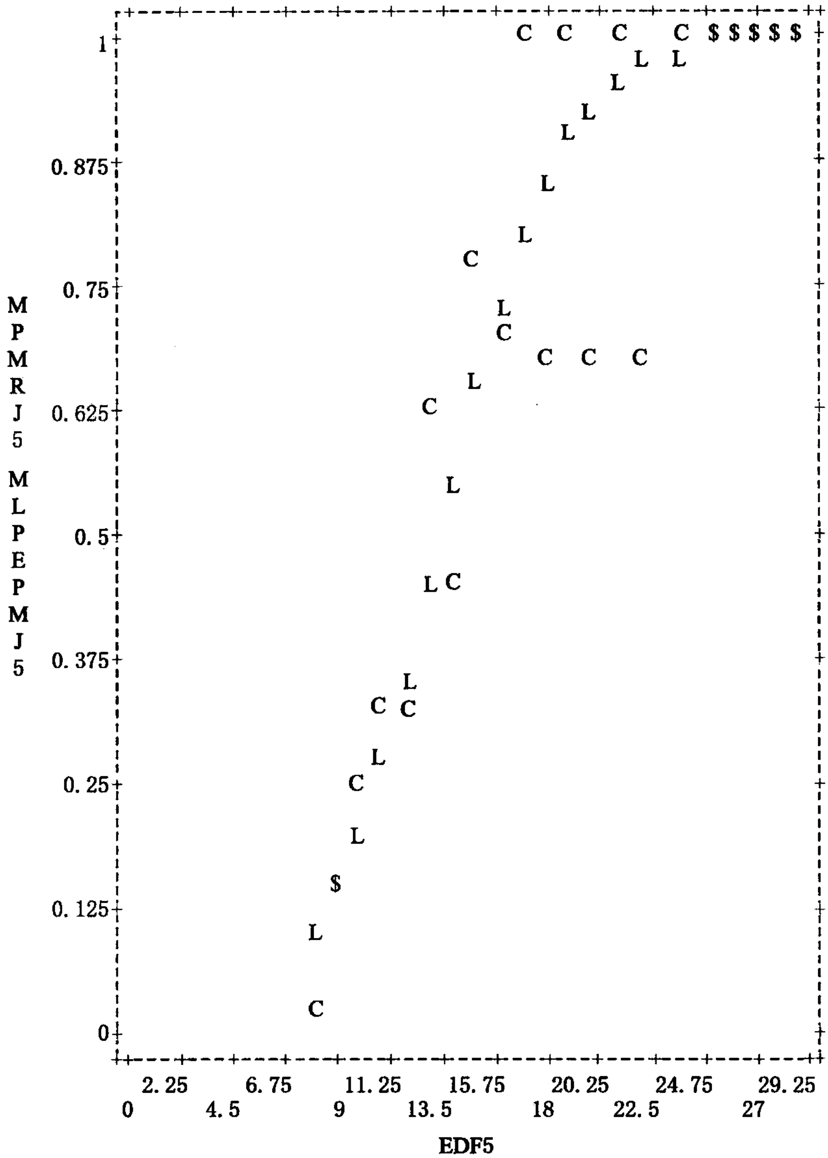
Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for yes
The Cut Value is 0.50
Symbols: n - no
y - yes
Each Symbol Represents 5 Cases.

图 1.4 大麻使用频率的二元 logistic 回归

部分 SPSS 的输出被删除但没有加入任何东西。图 1.4 中的大麻使用率的对数方程式是：



注：C：大麻普及率的观测平均值(MPMRJ5)对 EDF5 接触违法朋友。
L：logistic 回归预测的大麻普及率(MLPEPMRJ5)对 EDF5 接触违法朋友。
$：重复出现(线性回归预测和观测一致)。21 个个案。

图 1.5 条件概率：观测条件概率(C)与 logistic 回归预测的条件概率(L)

$$\text{logit}(\text{PMRJ5}) = -5.487 + 0.407(\text{EDF5})$$

图 1.4 展示了其他几种统计资料，这些将于下文再做讨论。注意，logistic 回归包括：(1)模型的拟合优度的统计概要（模型系数综合检测和模型概要）；(2)根据个案回答是否使用大麻，比较其观测值与预测值（或分类）；(3) logistic 回归参数(B)的估计值，与其他参数的统计资料（方程式中的变量）；(4)使用大麻的观测（是＝1，否＝0）与预测概率的图（观测和预测概率）。

图 1.5 分别指出 logistic 回归的预测和观测条件概率（或同等的条件平均数）。观测条件概率用字母“C”表示，预测条件概率用字母“L”表示。在图 1.2 中，线性回归的预测概率是一条直线，当 EDF5 大于 23.5 时，使用大麻的预测条件概率会大于 1。观测条件概率不像预测条件概率，会止于 1。在图 1.5 中，logistic 回归分析中的条件概率介于 0 至 1 之间，预测概率的曲线图形随着观测条件概率而变，这曲线类似图 1.3 的右半条。如图所示，利用 logistic 回归来预测因变量、观测和预测条件平均数的联系较紧密。

第 2 章 | 评估 logistic 回归模型的统计概要

当我们评估一个线性回归模型时，其标准通常有三部分。第一，整体模型是否可行？我们有没有信心认为所有自变量与因变量相关，而这种相关性是否是出乎预期的巧合或是随机的样本变异呢？如果相关，强度有多大？第二，如果整体模型是可行的，每个自变量的重要性如何？变量间的相关是否基于随机的样本变异？如果不是，每个自变量对因变量又有多大的预测力？自变量是强还是弱、好还是差？第三，模型形式是否正确？是否符合模型假设？在这一章，我们会先说明第一个问题——模型的整体合适度。第 3 章将讨论每个自变量的贡献，第 4 章将讨论检验模型的假设检验。

关于线性回归分析我们需要明白：(1)集合所有自变量的已知数值是否能更好地预测因变量。如果是，(2)所有自变量集合起来能多有效地解释因变量。对于 logistic 回归，我们感兴趣的是正确与错误预测的频率，以及模型能有效地减少预测误差。在线性回归中，当因变量是定距(interval)或定比变量(ratio scale)时，其预测值不完全等于观测值，这并不奇怪。在 logistic 回归中，因变量只含几个可能数值(通常只有两项)，比起预测值有多接近其观测值(0 或 1)(预测条件平均数等于预测条件概率)，我们有时更关注预测的准确度。

R^2，F 和误差平方和

在线性回归分析中，整体模型的评估是基于两个平方和。如果我们想把预测的误差平方和降到最低，同时，如果我们只知道因变量的数值(但不是这些数值属于的个案)，我们可以用 $\overline{Y}-Y$ 的平均值，作为所有个案的 Y 预测值，以减

少预测的误差平方和。基于这样的预测，误差平方和将会是 $\sum(Y_j-\overline{Y})^2$ 一总平方和(Total Sum of Squares, SST)。如果自变量能有效地预测 Y，则 $\hat{Y}_j$ 一回归方程式计算出来的 Y 预测值(Y 的条件平均数)将会比 $\overline{Y}$ 值好，其误差平方和 $\sum(Y_j-\hat{Y}_j)^2$ 将比 Y 平均值的误差平方和 $\sum(Y_j-\overline{Y})^2$ 小。$\sum(Y_j-\hat{Y}_j)^2$ 称为误差平方和(Error Sum of Squares, SSE)，是 OLS 所选取参数(β_1，β_2，…，β_k)算出的最小数值。第三个平方和，回归平方和(Regression Sum of Squares, SSR)是 SST 和 SSE 之差：SSR = SST − SSE。

在样本分析中，即使自变量 X 与 Y 并不相关，但如果用回归方程式去预测 Y_j 的数值，预测误差可能会比 $\overline{Y}$ 明显减少。这是因为抽样方差所形成的现象，样本值的随机波动有可能导致两个变量出现相关性，尽管它们实际是不相关的。多元 F 检验用 $\hat{Y}$ 去预测比 $\overline{Y}$ 好，是否是因为样本的随机变异？具体地说，多元 F 检验两个等同的假设：H_0：$R^2=0$ 和 H_0：$\beta_1=\beta_2=\cdots=\beta_k=0$。对于 OLS 线性回归，$F$ 比率(N 个个案和 K 个自变量)的计算方法如下：

$$F=(\mathrm{SSR}/k)/[\mathrm{SSE}/(N-k-1)]=(N-k-1)\mathrm{SSR}/(k)\mathrm{SSE}$$

F 比率的统计意义(p)指的是当原假设(null hypothesis)成立时，算出的 R^2 等于观测的 R^2，或算出的 β 系数等于观测的 β 系数的概率。如果 p 值很小(通常小于 0.05，也可以选其他 p 值)，我们拒绝原假设并得出自变量与因变量的相关并非出于偶然。如果 p 大于 0.05 的话，我们不能拒绝原假设，则结论是没有足够的证据去肯定模型已解释的方差不是基于样本的随机变异，这并非表示我们认为变量之间没有相关性，即使有相关性，我们也没有足够的信心去肯定它真的存在。

决定系数或 R^2，或“已解释方差”(explained variance)(已解释方差的比例)是一个实质意义的指标，即我们关心的变量的关系是否“足够大”或“足够强”。R^2 是误差减少的比例(proportional reduction in error)的统计指标，它量度的是相对于预测平均值 $\overline{Y}$，利用回归方程式能减少预测误差的比例(或乘以 100 变成百分比)。R^2 的数值介于 0(自变量完全没用)至 1(自变量能完美地预测每个 Y_j)。R^2 的计算方法如下：

$$R^2 = \text{SSR}/\text{SST} = (\text{SST} - \text{SSE})/\text{SST} = 1 - (\text{SSE}/\text{SST})$$

F 比率和R^2 可表达成另一个函数：$F = (R^2/k)/[(1-R^2)/(N-k-1)]$ 和 $R^2 = kF/(kF + N - k - 1)$。

在一个大的样本中，变量间的关系很可能出现统计显著性 ($p \leqslant 0.0001$)，但 R^2 却没有实质意义(例如，$R^2 \leqslant 0.005$)。如果自变量只能解释小于 0.5%的因变量方差，一般可忽略，尽管我们相当肯定所解释的方差并非出于随机的样本变异。小样本的变量关系很可能有实际的意义(例如，$R^2 \geqslant 0.4$)，但没有呈现统计显著性。虽然变量关系呈现中等强度(已解释方差是 0.40，或减少 40%的预测误差)，但是我们没有足够的个案数量去证明这个结果不是基于随机的样本变异。

拟合优度：GM，R_L^2 和对数似然

类似 logistic 回归模型的 F 和 R^2，误差平方和是线性回归模型选择参数的准则，而对数似然则是 logistic 回归模型选择参数的准则。在显示对数似然的资料时，统计软件通常把它乘以－2，原因我们稍后再谈。对数似然乘以－2 可简写为－2LL。对数似然是负数，－2LL 是正数，其数值越大，表示因变量预测越差。只含截距的 logistic 回归模型的－2LL 数值，在 SPSS LOGISTIC REGRESSION的输出中，就是模型系数综合检验表(ominibus tests of model coefficients)内的卡方统计加上模型概要表(model summary)的对数似然值的－2 倍(见图 1.4)。在 SPSS NOMREG 和 PLUM 中，这一数值已出现在模型拟合数据表(model fitting information)中，详情稍后再讨论。在 SAS 的输出中，也就是 SAS PROC LOGISTIC 的“只有截距”栏中的－2LOG。“只有截距”或首个－2LL，写成 D_0，表示没有任何自变量方程式的对数似然值的－2 倍，类似于线性回归模型的平方总和。对于二分因变量(编号为 0 或 1)，如果 $n_{Y=1}$ 是 $Y = 1$ 的个案数量，N 是个案的总数，$P(Y = 1) = n_{Y=1}/N$ 就是 $Y = 1$ 的概率，那么：

$$D_0 = -2\{n_{Y=1}\ln[P(Y=1)] + (N - n_{Y=1})\ln[1 - P(Y=1)]\}$$
$$= -2\{(n_{Y=1})\ln[P(Y=1) + (n_{Y=0})\ln[P(Y=0)]\}$$

logistic 回归模型的 -2LL，包括自变量和截距，就是 SPSS LOGISTIC REGRESSION 模型概要表中的 -2 对数似然值，也就是 SPSS NOMREG 和 PLUM 模型拟合数据表中最终模型的 -2 对数似然值，以及 SAS PROC LOGISTIC 的“截距和协变量”栏的 -2LOG L 统计称为全模型的 D_M。D_M 类似于线性回归中的误差平方和。logistic 回归分析中最类似线性回归的回归平方和就是 D_0 与 D_M 之间的差异 ($D_0 - D_M$)，叫做模型卡方(model chi-square)(在 SPSS LOGISTIC REGRSSION 的综合检验表中)，或最后模型的卡方(在 SPSS NOMREG 和 PLUM 的模型拟合数据表中)，或 -2LOG L(在 SAS PROC LOGISTIC 的“协变量卡方”栏中)。它被称为 G_M 或模型 χ^2。

在 logistic 回归(和其他线性模型)中，两个对数似然值之差乘以 -2，可解释为 χ^2 统计。如果它们来自两个不同的模型，而且其中一个嵌套在另一个模型中(McCullagh & Nelder, 1989)，即第一个模型包括第二个模型的部分但非全部的预测变量，同时又不包括第二个模型没有的预测变量。换言之，第一个模型的预测变量就是第二个模型的一个子集(subset)。D_M 可直接解释为第一个模型(只含截距)和第二个模型(包含截距和一个或多个预测变量)之间的差别。卡方统计 D_M 可检验 logistic 回归的原假设 ($\beta_1 = \beta_2 = \cdots = \beta_k = 0$)。如果 D_M 统计上有显著性 ($p < 0.05$)，我们可以拒绝原假设并得出结论，比起不包含任何自变量，包含变量的模型能对 $P(Y = h)$ 有更好的预测(h 是特定值，通常是二分因变量的 1)。因此，D_M 就像线性回归中的多元变量 F 检验(multivariate F test)和回归平方和。

麦卡拉和内尔德(McCullagh & Nelder, 1989)等人提到的“偏差行为”(deviance)(当讲到使用大麻的例子时，这一术语就很容易混淆，所以，后文会尽量避免)，D_M 总是用于测量拟合优度，是检测 logistic 回归模型未解释方差(unexplained variance)的统计意义的检验，类似于检验 OLS 回归模型的未解释方差的统计意义。打个比喻，G_M 是问这个杯子有多满(预测变量能多有效地预

测因变量)，而 D_M 是问这个杯子有多空。G_M 是比较只含截距模型与全模型(这一模型包括所有预测变量)，D_M 是比较全模型(full model)与饱和模型(saturated model)包括所有预测变量及可能的交互项(interaction)。旧版的 SPSS LOGISTIC REGRESSION 是假定 D_M 接近 χ^2 分布和不同程度的统计显著性，但将 D_M 作为 χ^2 统计出现的问题，就是用不同的方法去定义饱和模型，会导致不同的 D_M 值和不同的自由度(Simonoff，1998)。

西蒙诺夫(Simonoff，1988)简略地说明了两种方法，一种以个案为主(casewise approach)，这是 SPSS LOGISTIC REGRESSION 和 SAS PROC LOGISTIC 采用的方法，它将每个个案设定为相互独立的，自由度(degree of freedom)为 1。另一种方法是列联表法(contingency table approach)，它将每个预测变量值的组合或共变规则(covariate pattern)作为交叉列表，分析(crosstabulation)每一格，根据共变规则的数目(列联表的格数)而非个案的数目去计算自由度。无论哪一种方法，如果每格的个案数目太小或存在过多空格，在一般情况下，D_M 都不会呈 χ^2 分布，因此不适合用 χ^2 统计去检验拟合优度(McCullagh & Nelder，1989；Simonoff，1998)。如果有大量与共变规则的数目和每格中的个案数目相关的案例，就可能被定义为一个合适的饱和模型，同时可以计算偏离统计。偏离统计是呈 χ^2 分析，自由度是根据列联表方法修正得出的。这就是 SPSS NOMREG 和 PLUM 采用的，两者都可分析二分变量和含有多于两类的名义(nominal)或定序(ordinal)变量。NOMREG 和 PLUM 的拟合优度信息中含有皮尔逊(Pearson)和偏离 χ^2 统计，后者是基于－2 的对数似然值。

就个体层次的(casewise)数据而言，仍有可能去建立一个拟合优度指标。二分因变量最常用的是霍斯默和莱默苏(Hosmer & Lemeshow，1989)的拟合优度指标 $\hat{C}$，SPSS LOGISTIC REGRESSION 或 SAS PROC LOGISTIC 都有。霍斯默和莱默苏的拟合优度指标将数据压缩到小数点，并且基于对某兴趣特质预测的几率(例如成为吸食大麻的人)。另一些可能的拟合优度指标包括赤池信息量准则和施瓦兹准则(AIC 的修订版)，这两个在 SAS PROC LOGISTIC 中

都有显示。统计分数(例如 G_M)是模型自变量的综合效果的统计意义检验。博伦(Bollen)(1989)曾简单地介绍过赤池信息量准则和施瓦兹准则，这是用于比较模型的相关指针，多用于提供模型拟合的检验，我们可以比较拟合模型的赤池信息准则和施瓦兹准则，以及只有截距模型的 AIC 和施瓦兹准则，这比 G_M 能提供更多的数据。

对某些研究人员，特别是对对数线性模型或一般线性模型有一定认识，或是倾向于理论研究的人士，拟合优度将是一个重要的考虑。如果 logistic 回归模型的目标(单一因变量的预测)与应用焦点、线性与 logistic 回归的比喻都一致，很多时候都会集中讨论 G_M。

自变量与因变量多元关系的测量

前面已提到数个 logistic 回归中与线性回归 R^2 类同的比喻。哈格和米切尔(Hagle & Mitchell, 1992)、梅纳德(2000)、维尔和齐默尔曼(Veall & Zimmerman, 1996)的著作中有更详细的介绍。现在的焦点是 R^2，一般常见的软件如 SAS 和 SPSS 都会给出，决定系数的几个常见类别也会有比较。如果我们保留 logistic 回归的 $-2LL$ 统计量和线性回归解释平方和/误差平方和，SSR/SST 的直接模拟统计量就是似然比 R^2，$R_L^2 = G_M/(D_0) = G_M/(G_M + D_M)$ (McFadden, 1974; Agresti, 1990: 110—111; DeMaris, 1992: 53; Hosmer & Lemeshow, 1989: 148; Knoke & Burke, 1980: 41; Menard, 2000)。R_L^2 是量度 $-2LL$ 减少的比例(proportional reduction)或对数似然值绝对值减少的比率，$-2LL$ 或对数似然绝对值，最小的数值是用以选择模型参数——作为“变异”的量度(Nagelkerke, 1991)，虽不是完全与 OLS 回归中的变异一致，但也可做模拟。R_L^2 表示模型包含的自变量可以降低的方差，可以用 D_0 表示。这个变异介于 0 ($G_M = 0$ 的模型，$D_M = D_0$，表示自变量无法预测到因变量)和 1 ($G_M = D_0$ 的模型，$D_M = D_0$，表示自变量能完美地预测因变量)之间。R_L^2 可从 SPSS 的 NOMREG 和 PLUM 的输出结果中直接得到伪 R^2 表中的 McFadden R^2。奇怪的是，就发生比而言，SPSS LOGISTIC REGRESSION 并没有

输出比值(至少在撰写本文时尚如此),R_L^2 必须从 D_0(或 D_M)和 D_M 中自行计算出来[7]。

现在 SPSS 和 SAS 用的两个量度是:(1)每个观测改善数值的平均数的平方 $R_M^2 = 1-(L_0/L_M)^{2/N}$, L_0 是只含截距模型的对数似然函数,L_M 是包括所有预测变量的对数似然函数,N 是总个案数(Cox & Snell, 1989; Maddala, 1983:39—40);(2)调整后每个观测所改善的平均数的平方 R_N^2(Cragge & Uhler, 1970; Maddala, 1983:40; Nagelkerke, 1991)。未调整的量度不会有 1 的数值,即使模型是完美的;调整后利用除以 R_M^2 在某特定的因变量和特定最大可能值的量度允许有 1 的数值: $R_N^2 = [1-(L_0/L_M)^{2/N}]/[1-(L_0)^{2/N}] = R_M^2/(R_M^2$ 的最大可能值)。在 SPSS LOGISTIC REGRESION 中,R_M^2 和 R_N^2 分别是模型概要中的考克斯—斯内尔和 Nagelkerke R^2,或 SPSS NOMREG 和 PLUM 输出的 Pseudo-R^2 表中的考克斯—斯内尔和 Nagelkerke 的伪 R^2。在 SAS PROC LOGISTIC,它们只简单地写做 R^2 和调整后的 R^2。

一系列可替代的 R_L^2,包括奥尔德里发生比和纳尔逊(Aldrich & Nelson, 1984)在 logit 模型和 probit 模型中提出的伪 R^2 或偶然系数 R_C^2;沃尔德(Wald) R_W^2(Magee, 1990),以及麦凯尔维—扎沃纳(McKelvey & Zavoina, 1975) R_{MZ}^2。沿用本文的符号,如果 N 是个案数量,则 $R_C^2 = G_M/(G_M+N)$。沃尔德 $R_W^2 = W/(W+N)$,W 是多元沃尔德统计。麦凯尔维—扎沃纳的概率模型 $R_{MZ}^2 = s_{\hat{Y}}^2/(s_{\hat{Y}}^2+1)$(最早发展出来的)或 logit 或 logistic 回归模型的 $R_{MZ}^2 = s_{\hat{Y}}^2/(s_{\hat{Y}}^2+\pi^2/3)$,$s_{\hat{Y}}^2$是 $\hat{Y}$(Y 的预测值)的方差,1 和 $\pi^2/3$ 分别是标准正态分布(standard normal distribution)和 logistic 分布(logistic distribution)的标准差(standard deviation)。这些量度都有共同的特征,他们均不能有数值 1——完美模型。哈格和米切尔(1992)修正了奥尔德里发生比和纳尔逊的伪 R^2 可以由 0 至 1;从理论上讲,这个方法能同样应用在沃尔德和麦凯尔维—扎沃纳量度中。

哈格和米切尔也提出,修正后的 R_C^2 能对 OLS 回归提供一个好的估计值,维尔和齐默尔曼提出,麦凯尔维—扎沃纳 R_{MZ}^2也有同样的结果。当二分因变量

代表一个潜在间距的量表时，这个例子有数个不同的代替，包括利用线性概率模型的可能性（因为限制二分潜在间距量表的数值不太真实）；多元相关和加权最小平方估值应用于较为复杂的结构方程模型（Jöreskog & Sörbom, 1993）；R^2 量度因变量的观测值和预测值的相关强度（strength of association）。

R^2 是 OLS 线性回归分析常见的判定系数，在 logistic 回归的文献中相对比较少被关注（Agresti, 1990：111—112），其在 logistic 回归中的应用也颇受质疑。因为不同于 R_L^2、奥尔德里发生比和纳尔逊的伪 R^2，它不是基于选取模型参数的准则，而且如果二分因变量是一个潜在变量的量度指标的话，R^2 会提供一个偏离的解释方差估计。但 R^2 有不能完全为 R_L^2 所取代的优势，它可补充自变量与因变量间的关系。R^2 还有以下几项优势：第一，当预测观察值为某特定值时（代替预测观察概率，因变量等于该值），R^2 允许 logistic 回归分析模型与线性概率（linear probability）、方差分析（analysis of variance）和判别分析（discrimination analysis）模型做直接比较；第二，R^2 对计算标准 logistic 回归系数很有用；第三，现时的统计软件相对比较容易计算出来。

SPSS 和 SAS 是怎样计算 logistic 回归的 R^2 呢？假设因变量 Y，Y 的 logistic 回归模型预测值变量为 LPREDY，假设因变量 Y 在 LOGISTIC 回归模型中的预测值为 LPREDY，并在 SPSS 中进行设定。SPSS LOGISTICS REGRESSION (SAVE = PRED[LPREDY])，或在 SAS 中设定 SAS PROC LOGISTIC(OUTPT PRED)。接着，使用二元或多元回归步骤（例如 SPSS REGRESSION 或 SAS PROC REG）去计算 R^2。另外，可用方差分析去计算 η^2 或 η(SPSS MEANS 或 ANOVA；SAS PROC GLM 或 ANOVA)，因变量 Y 的观测值作为自变量，Y 预测变量 LPREDY 为因变量。因为这里只有两个变量（Y 是一个变量，Y 预测值是另一个变量），所以 $\eta^2 = R^2$，两者可互相变换。η^2 的角色可在因变量 Y 与其预测值之间互相变换（基于自变量值），虽然听起来有点怪，但是计算方法的确与判别分析的典型相关（canonical correlation）一致（Klecka, 1980）。

根据各量度的不同特性，笔者认为（Menard, 2000），R_L^2 最适合 logistic 回归[8]，首先也是最重要的是，相对于 R^2、R_W^2 和 R_{MZ}^2，R_L^2 在概念上最接近于

OLS 的 R^2，它反映最小的减少比例（$-2LL$；等同于最大的对数似然）。与那些依赖样本大小和对数似然或$-2LL$（R^2_M，R^2_N，R^2_C）不同，R^2_L 只依赖那些被放到最大或缩到最小的数值。第二，R^2_L 对基准比率（base rate）的敏感度不大，基准比率指的是个案特质的比率（例如，是不是吸食大麻者）。证据显示 R^2_M，R^2_N，R^2_C 和 R^2 均有不好的特质，他们会随着基准比率（哪个比较小，$n_{Y=1}/N$ 或 $n_{Y=0}/N$）由 0 增加至 0.5 而上升，所以有人开玩笑说，样本大小可当做决定系数来量度解释方差（Menard，2000：23）。第三，R^2_L 不像未调整的 R^2_W，R^2_C 和 R^2_{MZ}，它介于 0 至 1 之间，0 代表自变量完全没有预测能力，1 代表完美预测。第四，维尔和齐默尔曼（1996）提出，R^2_L 也能有效应用于多分类名义变量（polytomous nominal）和定序因变量，能与二分因变量媲美，不像以方差为本的量度 R^2_{MZ}和 R^2。

预测效率：λ_p，τ_p，ϕ_p和二项检验

除了关于拟合优度的统计，logistic 回归程序一般还包括不同的列联表，指出分析个案的因变量的预测数值和观测值。这些表类似 SPSS CROSSTABS 和 SAS PROC FREQ 的列联表。很多时候我们比较在意模型准确预测的概率 $P(Y_j=1)$，但有时我们可能对准确的分组预测更有兴趣，因此分类表格与模型整体拟合同样更受关注。基于 logistic 回归或相关方法如判别分析，现在还没有关于如何量度个案分类的观测与预测结果之间的关系的一致说法。但是，几个方法都能轻易地得到综合概要的列联表，最好选用一些 logistic 回归软件，过程会涉及误差改变比率（proportional change in error）的计算。

$$\text{预测效率}=\frac{(\text{没有模型的误差})-(\text{有模型的误差})}{(\text{没有模型的误差})} \qquad [2.1]$$

这就是误差改变比率方程。如果模型能更有效地预测因变量，就会和误差比率减少一样。但是在某些情况下，可能模型比几率更差。此时，预测效率会出现负数，误差按比例增加。简单来说，模型的误差指的是自变量预测值错误

的个案数量。没有模型的误差与这三个指标不同，同时要看是否为预测、分类或选择模型。

预测、分类和选择模型

预测模型并不预先限制各种特定行为或特质的个案数目或比例。原则上，有可能(但不是必然)被预测为"正"(具有某特质，例如"成功")或"负"(没有某特质，例如"失败")的个案的数量与观测到的"正"与"负"的数量是相同的。也就是说，没有限制预测或观测频率的边际分布(marginal distribution，正或负，每一类的个案数量或比率)要相等或不等。更有甚者，全部个案都被预测到同一类别，即样本或全部人口都是同质的(homogeneous)。实际上，当所有组别的特质都相同时(全锁住或全放走他们)，预测模型是合适的。

分类模型与预测模型很相似，但它多做了一个假设，就是个案必须是异质的(heterogeneous)。因此，分类模型的评估必须加强限制，即模型把个案分到每一类的数目要与实际观测到的数量相同。每类观测个案的数量或比例必须与其预测的相同。如果一个模型不能符合这个准则，它就不是分类模型。完全同质是分类模型不能接受的。以现实情况来说，假设异质，而且所有组别不可能是相同的，那么分类模型就是合适的。

选择模型(Wiggins，1973)所关注的是"接受"或"拒绝"个案，准则是指是否满意该组别的某些要求，同时符合该组可接受的最低要求、最大限量、特定的个案数量。在选择模型中，观测到成功(或基准比率)的个案比率可能或不可能等于该组接纳或已选择的个案比率(选择比率)。例如，某公司从 200 名职位申请者中聘请 20 名员工，选择比率是 20/200 = 0.10(10%)，无论基准比率(成功被聘用的观测概率)是 5%或 20%，是选择比率的一半或两倍。logistic 回归软件的分类表自然成为预测或分类模型。虽然以前他们有可能是选择模型，但是必须转换(除非在巧合的情况下，选择比率刚好等于基数率)，这样，选中的个案数目才是正确的。

列联表作为预测效率的指标：常见的相关衡量

在不同的量度预测效率的指标中包括数个用以分析列联表相关的量度：ϕ，古德曼和古鲁斯克(Goodman & Kruskal)γ、κ，列联系数，皮尔逊 r 和发生比数比(Farrington & Loeber，1989；Mieczkowski，1990；Ohlin & Duncan，1949)。使用一般列联表的相关衡量去分析 2×2 或更大的预测表格的问题在于：区分(1)自变量 X 与因变量 Y 关系的强度；(2)预测组别 $E(Y_j)$ 与观测组别 Y_j 关系的强度。图 2.1 说明了两者的区别。图 2.1 中的表格 A 代表常见 2×2 表格的格数和边际频率，表格 B 代表种族与政治取向的假定关系，表格 C 说明预测与观测政治取向的假定关系。

虽然表格 B 和 C 的数字是完全相同的，但是推论却不同。在表格 B 中，种族的数据可令预测政治取向误差比率减少(Proportional Reduction in Error，PRE)约 0.20(按古德曼—古鲁斯克的 λ)或 0.04(按古德曼—古鲁斯克的 τ)。在表格 C 中，当我们预测的与假定模型预测的相反时，PRE 一样。事实上，用模型预测政治取向比几率更差，原因有可能是数据的偏斜，或用其他数据发展出来的预测模型去预测。如果每个个案的分类都错了，表格 C 中的 λ 和 τ 都是 1；因此，完全正确与完全错误便没有区别。皮尔逊 r 和它等同 2×2 表格的肯德尔(Kendall)τ 和 ψ 计算如下：

$$\phi=(ad-bc)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$$

会显示负符号的分类，它的平方便是 PRE。

但是，对没有序列类别的大表格，皮尔逊 r 和肯德尔 τ 都不能用，而 ϕ 变成克拉默 V，也不能再有 PRE 的解释。发生比数之比可用在 2×2 表格上，但对大表格就需要计算两个或两个以上的发生比数之比，而且不再提供正确预测的单一综合概要。总体来说，列联表与预测表的相关常见衡量不再对预测准确度的估量提供直接而整体的答案。皮尔逊 r 和 r^2 或 ψ 和 ψ^2 是 2×2 表格合理的指针，但对于较大的表格就不适用。

表格 A:预测表的标准格式

		Y 的预测值		
		正面(成功)	负面(失败)	
Y 的观测值	正面(成功)	a	b	$a+b$
	负面(失败)	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

表格 B:种族与政治取向

		X:种族		
		欧裔	非欧裔	
政治取向	保守	20	30	50
	自由	30	20	50
		50	50	100

表格 C:预测与观测政治取向

		预测的政治取向		
		保守	自由	
观测的政治取向	保守	20	30	50
	自由	30	20	50
		50	50	100

注:表格 B 和表格 C 的古德曼—古鲁斯克的 $\lambda=0.20$;古德曼—古鲁斯克的 $\tau=\phi^2=r^2=0.04$。

图 2.1 关联性与预测

λ_p, τ_p 和 ϕ_p

方程式 2.1 提供了预测效率指标的基本方式。模型误差就是模型错误分类的个案数量,它类似于误差平方和。没有模型的总误差类似于总平方和。这要根据我们是否使用预测模型、分类模型或选择模型而定,对预测模型,最类似线性回归(间距因变量)的方法就是将因变量的模式作为所有个案的预测值(类似用线性回归的平均数)。定义没有模型的误差的方法与名义变量的列联表的古德曼—古鲁斯克 λ 一样,是奥林(Ohlin)和邓肯(Duncan)(1949)首次建议的一个指标。因为与古德曼—古鲁斯克 λ 很相似,它把这一指标称为 λ_p(lambda-p), p 指的是运用预表格。

当 Lambda-p 为正数时，它就像 R^2 一样，是一个关于 PRE 的衡量，但如果模型比预测模式差，λ_p 有可能是负数，显示误差按比例增加。λ_p 的可能值依赖于边际分析。一般而言，所有表格(N 个个案)的 λ_p 可能值的范围是由 $1-N$ 至 1。

在分类模型中，没有模型的期望误差定义如下：

$$没有模型的误差 = \sum_{i=1}^{N} f_i[(N-f_i)/N]$$

N 是样本数量，f_i 是观测类别 i 的个案数目，这与古德曼—古鲁斯克 τ 的个案数目差不多。克莱卡(Klecka, 1980)提出了用以判别分析的没有模型的衡量指针。与 λ_p 平行的克莱卡指标是 τ_p(tau-p)或预测表格的 τ。

正如 λ_p 一样，τ_p 是量度变化的，但 τ_p 需要把个案分到不同的组别或类别中，不能放在同一类别中，即使在没有模型的误差估计中。实际上，τ_p 能调整分类基数率的误差期望数量。预测的准确度因此可算其次，以异质假设为准。如 λ_p，τ_p 等于 1 表示所有个案分类全部准确，τ_p 是负值表示预测模型比期望个案预测分类(基于观测边际分布)更差。τ_p 的第二个性质，二分因变量的 $\tau_p \geqq \lambda_p$，因为没有模型误差的 τ_p 的数目会等于或大于因为没有模型误差 λ_p 的数目。在相同边际分布的表格中，τ_p 介于 -1 至 1 之间，但当边际分布不相等时，最大的 τ_p 小于 1。最差的可能性是，边际分布极为偏斜和不一致。τ_p 的最小值等于 $1-N^2/2(N-1)$。在不同的边际分布中，τ_p 的范围不是固定的，介于 1 至 2 之间。负 τ_p 和大绝对值的模型会有比较窄的范围。

同时还有可能建立量度选择模式中预测准确度的误差比例变化。有模型误差是 $b+c$，类似于 λ_p 和 τ_p。有模型误差是基于基数率 $B=(a+b)/N$ 和选择率 $S=(a+c)/N$ 的。当知道 B, S 和 N 的值时，我们会知道格 a 的预测值(图 2.1 的表格 A)：$E(a)=\mathrm{BSN}$。因为一个 2×2 的预测表的自由度只有 1，只要 a 的预测值已知，边际分布、其他格的预测值也就可知，那么就与要计算 χ^2 统计的预测值一致。期望误差是 $E(b+c)=[(a+b)(b+d)/N+(c+d)(a+c)/N]$。将这些数值放到 PRE 算式内，我们会得出误差量度的比例变化[9]：

$$\phi_p = \frac{(a+b)(b+d)/N+(c+d)(a+c)/N-(b+c)}{(a+b)(b+d)/N+(c+d)(a+c)/N}$$

$$= \frac{(a+b)(b+d)+(c+d)(a+c)-N(b+c)}{(a+b)(b+d)+(c+d)(a+c)}$$

$$= \frac{ad-bc}{0.5[(a+b)(b+d)+(c+d)(a+c)]}$$

相同边际分布的表格，ϕ_p 最大值是 1。一般来说，它介乎于 -1 至 1 之间，但实际最大值、最小值和范围就得看边际分布。只要没有模型的误差就可以算是格 b 和 c 的期望频率之和(图 2.1 的表 A)，有模型的误差就可以算是格 b 和 c 的观测频率之和，那么，ϕ_p 就可延伸到比 2×2 大的表格上，而且它仍可解释成误差的比例改变。2×2 表格可以显示出 $|\phi_p|\leqslant|\phi|$，这时 ϕ_p 与 ϕ 和皮尔逊 r (分子一样，$ad-bc$) 都有同样的标志。当所有个案预测都正确时，$\phi_p=1$；否则，即使在特定边际，个案的正确分类已达到最大的值，$\phi_p<1$ [10]。

λ_p，τ_p 和 ϕ_p 的统计意义

Lambda-p，tau-p，phi-p 统计意义的量度类似于 R^2。设定 $N=$ 总样本大小，$P_e=$ (没有模型的误差)$/N$，$p_e=$ (有模型的误差)$/N$。二项统计 d 的计算方法如下：

$$d=(P_e-p_e)/\sqrt{P_e(1-P_e)/N}$$

d 接近正态分布(Bulmer，1979)[11]。注意，我们所比较的不是每个类别的个案比例，而是模型将正确或不正确的个案分类的比例。这种检验与预测、分类和选择模型的 λ_p，τ_p，ϕ_p 相同，不同的只是定义没有模型的误差。

在所提及的统计意义的检验中，观测分类的数值是已知的，检验显示用模型预测的错误(基于模型，因此会变异)比例与没有用模型预测的错误比例(只基于边际分布，不是模型，因此是固定的)有没有明显区别？二项检验的形成，明确地利用模型所产生误差的期望值为准则，这比两比例差距的二项检验好。两比例差距假设：有模型和没有模型的误差的比例是基于不同的样本(不相等的样本数量)。问题是，当我们比较 logistic 回归产生的分类表中的观测与预测

分类，或表格中的期望与真实的误差时，有一个条件明显不符合。但是如果检验两个不同的预测模型的整体预测准确度（正确预测的百分比）有没有明显分别，两比例差距的二项检验是很有用的。在这种情况下，我们需要分开去检验两个或任意一个模型是否比复制个案观测分类的几率好。

预测效率的其他指标

马黛拉（Maddala，1983:76—77）曾对计量经济学文献提及的三个预测效率指标进行评论。他摒弃其中一个，因为它不能分辨完全正确与完全错误的模型，这与其他通用的列联表的相关量度有相同之处。第二个指标被认为是“最好”或“次好”的，它与拟合优度很接近。第三个指标介于−1至0.50，取决于边际分布，其值与 ϕ 和皮尔逊 r 相似。因为它没有 PRE 解释，所以并不比 ϕ 或皮尔逊 r 好。

洛伯（Loeber）等人（1983）建议过一个方法，叫做相对机会提高率（Relative Improvement Over Chance，RIOC）。虽然洛伯和他的同事把 RIOC 应用在预测分析和分类表中，但是它能修正基数率与选择率的不同，因此，比起分类或预测模型，它更适合用于选择模型。RIOC 与系数 ϕ'（边际分布 ϕ 的修正系数）一致。与 ϕ 不同，ϕ' 没有 PRE 解释。这一数值介于−1（完全不正确的预测，如果图 2.1 中表格 A 的格 b 和 c 都不是零）和 1 之间。如果格 b 或 c 任何一格等于 0，即包含错误的预测，那无论正确个案分类的比例有多小，RIOC＝1，这与 Yule Q（列联表相关衡量）之一有同样的问题。即使有超过 90% 的个案被错误地分类，RIOC 也有可能等于 1[12]。

预测效率指标的比例

梅纳德（2000）提出一些研究证据，控制基数率，λ_p 和 ϕ_p 与基数率高度相关，但 τ_p 没有。索德斯特龙（Soderstrom）等人（1997）先控制模型的基数率、样本大小和预测变量的信度，然后比较 λ_p，τ_p，ϕ_p，RIOC 和正确百分比。根据 Monte Carlo 模拟，他们发现在含有连续预测变量的模型中，ϕ_p 最不受基数率的影响，接着是 τ_p 和 λ_p。只含有二分预测变量的模型，对基数率的改变，τ_p 没有

λ_p 敏感，RIOC 通常计算不出来，原因前文已谈过。索德斯特龙总结，在很多情况下，ϕ_p 和 τ_p 是最合适的预测效率。总的来说，这个结论深化了这个观点，即选择指针需要与模型的性质一致，特别是 τ_p 最适合分类模型，ϕ_p 最适合选择模型。这些指针可应用在 logistic 回归预测表和过程产生的预测表上，而且并不限于二分变量；只要预测表能分辨正确预测与不正确预测，这些指标都适用。不幸的是，现在通用的 logistic 回归软件，例如 SPSS LOGISTIC REGRESSION 或 SAS PROC LOGISTIC，都不会自动出现以上预测效率指标。

举例：评估 logistic 回归模式的充足性

研究人员通常对拟合优度比较感兴趣，这是预测效率缺乏共识的一个原因。

在理论验证上，拟合优度比分类准确度更为重要。我在文中介绍了预测准确度，但相对于拟合优度评估的发展和它的重要性来说，这方面的发展相对比较慢。通常，拟合优度和预测准确度这两个方法会得出一致的结果，但也完全有可能得出这样的结果：有完好的模型，却无法正确地分类。

图 2.2 和图 2.3 说明了拟合优度和预测效率如何导致非常不同的结论。在图 2.2 中，单一因变量 TRUE 是仿真的数据，另有单一预测变量 P1。标准的输出包括 R_L^2，R^2，λ_p，τ_p。图 2.2 分析了 40 个个案，模型拟合得非常好。$G_M =$ 模型卡方 $= 20.123$，在统计上是显著的 ($p = 0.0000$)，因此我们拒绝原假设，自变量 P1 与因变量 TRUE 不相关。$R_L^2 = 0.363$，说明 TRUE 和 P1 有相当程度的相关。二项 d 与 λ_p，τ_p 相同（两者都是 50% 的期望误差）：$d = 5.060$，有显著性 $p = 0.000$。λ_p 和 τ_p 都等于 0.80，显示我们可根据自变量把个案分到因变量的不同类别，而且准确度也同样高，这在分类表中都已反映出来。整体来说，预测准确度比模型预测概率 $P(Y_j = 1)$ 的能力更高。图 2.2 下方的观测组与预测概率的图像，显示了预测概率有时非常高，有时接近 0.5——分类 ($Y = 1$ 或 $Y = 0$) 的分线。预测准确度非常高，尤其有些个案分到 $Y = 1$ 的预测概率接近 0.5。

Classification Table(a)

			Predicted		
			TRUE		Percentage Correct
	Observed		0	1	
Step 0	TRUE	0	18	2	90.0
		1	2	18	90.0
	Overall Percentage				90.0

a The cut value is 0.500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20.123	1	0.000
	Block	20.123	1	0.000
	Model	20.123	1	0.000

R_L^2 = 0.363
R^2 = 0.408
Tau-p = 0.80
Lambda-p = 0.80

Model Summary

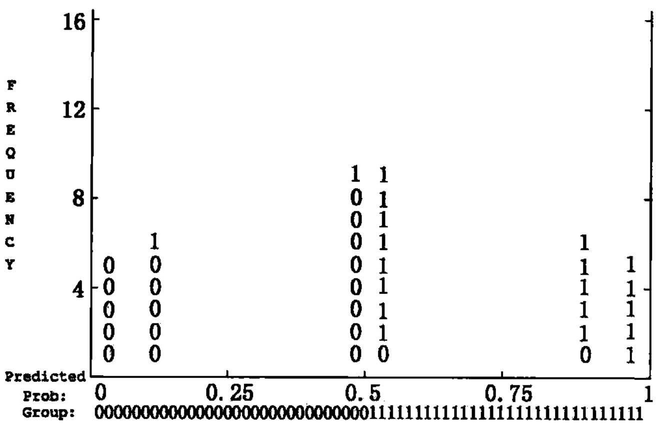
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35.329	0.395	0.527

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P1	8.203	2.894	8.035	1	0.005	3653.109
	Constant	-4.102	1.505	7.432	1	0.006	0.017

a Variable(s) entered on step 1: P1.

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for 1
The Cut Value is 0.50
Symbols: 0 - 0
1 - 1
Each Symbol Represents 1 Case.

图 2.2 假定"好的预测"的 logistic 回归输出(分类的预测概率)

在图 2.3 中,G_M 也有显著性,R_L^2 和 R^2 都显示因变量 TRUE 与新预测变量 P2 紧密相关。但 τ_p 和 λ_p 却显示观测和预测分类的相关性非常低,二项检验 $d = 0.632$,统计意义 $p = 0.264$ (单尾),反映将个案分到因变量与自变量的数

Classification Table(a,b)

	Observed		Predicted: TRUE 0	Predicted: TRUE 1	Predicted: Percentage Correct
Step 0	TRUE	0	11	9	0.0
		1	9	11	100.0
	Overall Percentage				50.0

a Constant is included in the model.
b The cut value is 0.500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	17.974	1	0.000
	Block	17.974	1	0.000
	Model	17.974	1	0.000

R_L^2 = 0.324
R^2 = 0.356
Tau-p = 0.10
Lambda-p = 0.10

Model Summary

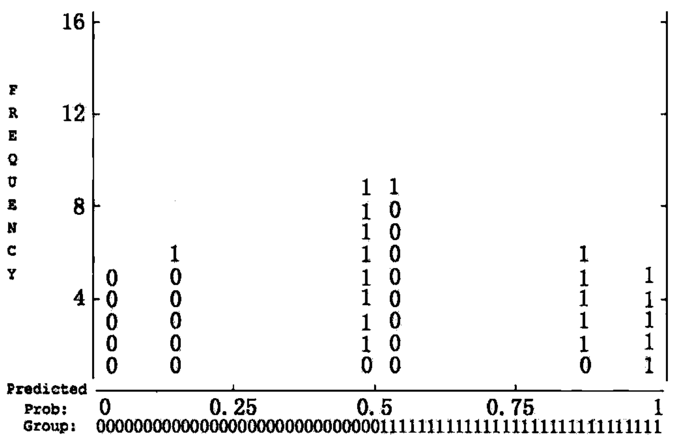
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	37.477	0.362	0.483

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P2	7.193	2.497	8.299	1	0.004	1329.985
	Constant	-3.596	1.311	7.527	1	0.006	0.027

a Variable(s) entered on step 1: P2.

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for 1
The Cut Value is 0.50
Symbols: 0 - 0
1 - 1
Each Symbol Represents 1 Case.

图 2.3 假设差的预测的 logistic 回归输出

值无关。原因显示在观测组与预测概率图中。当预测概率接近 0.5 时，不再是正确的预测，反而接近完全错误。40 个个案中的 26 个的 P1 与 P2 预测概率是一样的，但其余的 14 个，预测概率有 0.02 的改变，或从 0.49 至 0.51，或从 0.51

至0.49。这些对整体的拟合优度的R_L^2，R^2，η^2和G_M影响不大，但对预测效率的指标就有莫大的影响。

图1.4在第1章中讨论过，它显示了二元logistic回归分析的结果，接触违法朋友(EDF5)与大麻使用频率(PMRJ5)的关系。从图1.4的结果来看，根据G_M(意义 = 0.0000)，我们拒绝原假设，即EDF5与PMRJ5无关。注意：霍斯默和莱默苏拟合优度没有显著性，显示只含预测变量EDF5的模型与数据拟合度很高。预测表格显示预测准确度也不错，但我们需要计算λ_p和τ_p，通过数量化的衡量分析到底模型能多好地把个案分类。模型的预测值可储存成一个新变量，LPEPMRJ5(EDF5通过logistic回归预测PMRJ5的新变量)。这可以让我们利用另一个方差分析或二元回归去计算R^2。

根据图1.4的数据，$G_M = 85.359$(综合检验的模型卡方)和$D_M = 213.947$(模型综合概要的-2似然)，$R_L^2 = G_M/(G_m + D_M) = (85.359)/(85.359 + 213.947) = 0.285$。Lambda-p等于较小观测类别的个案数目($Y = 1$：$37+44 = 81$)减去模型错误分类的个案数目($37 + 14 = 51$)，除以较小类别的个案数目，因此，$\lambda_p = (81-51)/81 = 0.370$，相当程度地减少了预测误差。Tau-p的计算相对比较复杂。首先找出Y观测值的每个分类的个案总和：$Y = 0$，$n_{Y=0} = 150$；$Y = 1$，$n_{Y=1} = 81$。二分因变量的误差期望值是两个总数相乘，除以个案总数(231)并乘以2，因为我们期望二分变量的每类误差数目是相同的，即$(2)(150)(81)/231 = 105.2$。Tau-p是误差期望值减去误差的真实数目(51)，除以误差期望值：$\tau_p = (105.2-51)/105.2 = 0.515$。显示模型能把个案分为不使用或使用大麻类别的分类误差减少了至少一半。

我们可以利用λ_p和τ_p的期望误差去计算二项检验d统计。λ_p的误差期望数量是81，相应的比例是$81/231 = 0.351$，误差观测数是51，相应比例$51/231 = 0.221$。因此，$d = (P_e - p_e)/\sqrt{P_e(1-P_e)/N} = (0.351 - 0.221)/\sqrt{(0.351)(0.649)/231} = 4.140$

具显著性$p = 0.000$。τ_p不详细计算，$d = 6.996$，具显著性$p = 0.000$。最后，计算R^2或η^2，一定要用二元回归或方差分析的结果。我们可以比较图1.4

的 logistic 回归结果与图 1.1C 部分的线性回归的结果(同样的变量),logistic 回归模型的 PMRJ5 已解释方差 ($R^2 = 0.34$) 比线性回归稍微高点 ($R^2 = 0.32$,见图 1.1C)[13]。虽然线性回归是通过缩小平方误差总和以算出最大的 R^2,而 logistic 回归不是,但结果显示,logistic 回归更配合数据,这说明预测变量与因变量有相当程度的相关,而且分类也不错。

总结:评估 logistic 回归模型

在线性回归中,我们利用 F 统计和 R^2 去检验因变量和自变量相关性的统计显著性和实际意义。两者均基于平方总和 SST 与误差平方和 SSE。在 logistic 回归中,如果我们主要关注模型与数据是否拟合(例如用于理论建构与验证),那么我们用 G_M 和 R_L^2,基于 $-2LL$ 去检验统计和实际意义。如果我们较少关注整体模型的拟合度,而更关心模型预测的准确性,那么我们会用二项检验 d 和三个预测效率(λ_p, τ_p 和 ϕ_p)去检验模型的统计和实际意义。因为 logistic 回归软件没有提供 λ_p, τ_p 和 ϕ_p 三者的数据,所以必须另做计算。

第 3 章 | 解释 logistic 回归系数

在线性回归分析中，我们评估每个自变量对模型的贡献，是通过检验它的统计显著性来估计其对因变量的实际意义。统计显著性的评估是利用 F 或 t 检验产生的概率(p)，我们发现样本越大，自变量和因变量的关系越大，即使它们实际上不相关也是如此。实际意义有几种评估方法。我们可以检验非标准化回归系数，看看因变量的改变是否与自变量的改变相关，而且这种改变是否足够大(这里用“相关”比任何暗示因果关系的词汇好，虽然变量间关系有预测性，但可能本质上不是因果关系)。要应用该检验，我们必须事先知道改变要有多大，检验才会被关注，以及改变要多小，才可被忽略。非标准化回归系数对评估一个变量对另一个变量的实际影响和比较相同变量在不同样本中的影响是非常有用的。

另外，如果没有清晰的标准去说明到底多大才算大(有些变量的量度单位又不像尺、磅、元这种惯用量度单位，而是用量表数字，如接触违法朋友 EDF5)，我们就可以参看标准回归系数，它会显示随着自变量的标准偏差改变 1 个单位，因变量的标准偏差会改变多少。标准系数特别适用于理论验证和在相同样本中比较不同变量的影响。

逐步法(stepwise)用于评估回归方程中变量的贡献，尤其当我们检验回归方程的非线性(例如包括平方项)或非相加性(包括交互项)时。是否加入非线性或非相加的变量，这取决于解释方差 R^2 改变的大小和统计显著性。当我们关注理论发展多于理论验证时，逐步方法就是探索式分析。这种研究通常应用以上现象研究的前期阶段，这时现象的相关性理论或知识还没发展好。模型中逐步包含或删减变量所用的准则或方法通常与理论验证的方法差不多或稍宽

松些。

logistic 回归分析的统计显著性

自变量对解释因变量的贡献的统计显著性有几种评估方法。其中最好、最准确的是似然比检验(likelihood ratio test)。似然比检验是针对有变量和没有变量的 logistic 回归模型的一种方法。似然比检验统计量等于有变量模型 G_M 减去没有变量模型 G_M。此结果(当检验 X_1 时,称之为 G_1;检验 X_2 时,称之为 G_2;检验 X_k 时,称之为 G_k)呈 χ^2 分布,自由度等于有 X 模型的自由度减去没有 X 模型的自由度。例如,G_{M1} 代表模型中有 X_k 的 χ^2,G_{M2} 代表模型中没有 X_k 的 χ^2,$G_k = G_{M1} - G_{M2}$,如果 X 是连续、定距、定比变量,那 G_k 的自由度是 1。

唯一的缺点就是,似然比检验的计算比其他方法所需的时间长。如果主计算机需要付费,这将是一个严重的问题,但现在很多人使用较快的个人计算机,除非样本很大,一般不成问题。统计软件通常使用比似然比检验较易计算的方法——沃尔德统计,用以检验个别系数的统计显著性。如图 1.4,沃尔德统计排在系数(B)和标准误差(S. E.)之后。若个案的分布是不对称的如卡方分布,沃尔德统计的计算是 $W_k^2 = [b_k/(b_k$ 的 S. E. $)]^2$,若为标准正态分布,则沃尔德统计的计算是 $W_k = [b_k/(b_k$ 的 S. E. $)]$ (Hosmer & Lemeshow, 1989:31; SAS, 1989:1097; SPSS, 1991:140—141)。同时,沃尔德统计的算式与线性回归系数的 t 比率平行。沃尔德统计的缺点就是当 b 变大,其估计标准误差会增大,因此,若原假设是错误的,也会令我们无法拒绝原假设[14]。

图 3.1 是因变量与四个预测变量的 SPSS 输出。大麻使用频率 PMRJ5 仍是因变量,接触违法朋友 EDF5 仍是自变量,BELIEF4 被用来量度被访者认为 EDF5 所提到的每种非法行为的错误程度如何(完全错、有点错、没有错),包括侵犯他人、偷窃、贩卖毒品等。响应者先回答 BELIEF4,然后再回答是否使用。性别 SEX 的编号 0 是女性,1 是男性。种族 ETHN 的编号已在图 3.1 的分类变量编号表中显示。对应 ETHN 的第一个类别是非裔美国人(黑人),第二个类别

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	227	88.3
	Missing Cases	30	11.7
	Total	257	100.0
Unselected Cases		0	0.0
Total		257	100.0

a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0.00 no	0
1.00 yes	1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	108.257	5	0.000
	Block	108.257	5	0.000
	Model	108.257	5	0.000

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)
ETHN	1 white	175	0.000	0.000	
	2 black	37	1.000	0.000	
	3 other	15	0.000	1.000	
SEX	1 FEMALE	117	0.000		
	2 MALE	110	1.000		

Classification Table(a)

	Observed		Predicted PMRJ5 no	yes	Percentage Correct
Step 1	PMRJ5	no	134	13	91.2
		yes	28	52	65.0
	Overall Percentage				81.9

a The cut value is 0.500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDF5	0.407	0.069	34.341	1	0.000	1.502
	BELIEF4	-0.118	0.060	3.903	1	0.048	0.889
	SEX(1)	-1.514	0.405	14.008	1	0.000	0.220
	ETHN			1.190	2	0.552	
	ETHN(1)	0.245	0.508	0.232	1	0.630	1.277
	ETHN(2)	0.772	0.745	1.074	1	0.300	2.163
	Constant	-1.749	2.028	0.744	1	0.388	0.174

a Variable(s) entered on step 1: EDF5, BELIEF4, SEX, ETHN.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	186.359	0.379	0.522

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.754	8	0.363

Contingency Table for Hosmer and Lemeshow Test

		PMRJ5 = .00 no		PMRJ5 = 1.00 yes		Total
		Observed	Expected	Observed	Expected	
Step 1	1	24	23.306	0	0.694	24
	2	21	22.577	3	1.423	24
	3	24	22.379	1	2.621	25
	4	22	20.340	2	3.660	24
	5	18	18.097	5	4.903	23
	6	13	16.129	10	6.871	23
	7	14	13.079	9	9.921	23
	8	7	7.684	16	15.316	23
	9	3	3.129	20	19.871	23
	10	1	0.280	14	14.720	15

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	EDF5	-120.954	55.549	1	0.000
	BELIEF4	-95.222	4.085	1	0.043
	SEX	-101.185	16.011	1	0.000
	ETHN	-93.758	1.158	2	0.561

```
                         *** OUTPUT FROM SPSS MEANS ***

Summaries of   LRPRED13   Predicted Value By levels of PMRJ5

   Value  Label                 Mean      Std Dev   Sum of Sq    Cases

    0.00  no                 0.2002317  0.1958068  5.5976854     147
    1.00  yes                0.6320872  0.2965058  6.9453396      80
                            ---------------------------------------
Within Groups Total          0.3524275  0.2361076 12.5430250     227

                    Analysis of Variance

                          Sum of                 Mean
Source                    Squares      D.F.      Square        F       Sig.

Between Groups             9.6618        1       9.6618   173.3160   0.0000
Within Groups             12.5430      225       0.0557

                     Eta = 0.6596    Eta Squared = 0.4351
```

图 3.1　SPSS logistic 回归的输出结果

是除了非西班牙裔或非裔美国人之外的种族(其他)。在栏中标为(1)、(2)或(3),当个案落在该种族类别时,其相应变量取值等于 1。

因此,非裔美国人(黑人)就是第一个 ETHN 变量乘以 1,其他种族乘以 0;其他种族就是第二个变量乘以 1,黑人乘以 0。这是 logistic 回归应用的虚拟变量或设计变量,代表单一分类变量的例子,这种方法与线性回归是一致的(Hardy, 1993; Lewis-Beck, 1980)。

在图 3.1 中,方程式内的标签变量包括每个变量的 logistic 回归系数、标

准误差、沃尔德统计(W_k^2)、自由度(df)和沃尔德统计意义。图 3.1 显示，EDF5、BELIEF4 和性别对 PMRJ5 都有显著性，但 ETHN 无论是整体的还是分开的，虚拟变量对 PMRJ5 或截距统计上都没有显著性。霍斯默和莱默苏(列联表的霍斯默和莱默苏检验)显示模型拟合度高。在图 3.1 底部，SPSS MEANS 前面的 Model if Term Removed 标题下，显示出模型中各变量的似然比统计量[15]。实际上，沃尔德和似然比统计的结论、显著水平都很相似。之前提到过，考克斯—斯内尔和 Nagalkerke R^2 是量度有问题的量表，这里就不讨论了。

图 3.2 显示用 SPSS NOMREG 重做的输出结果。NOMREG 的输出结果较为简洁，与 LOGISTIC REGRESSION 显示同样的信息。模型 $\chi^2(G_M)$是一样的，最后的$-2LL(D_M)$也一样(比较图 3.1 的模型综合表和图 3.2 的模型拟合数据表)，只有截距或最初的$-2LL(D_0)$也是如此。两个表格的考克斯--斯内尔和 Nagelkerke 也是一样，但 NOMREG 还提供 McFadden 伪 R 平方(R_L^2)。图 3.2 的似然比检验与图 3.1 删除变量后的模型表提供了相同的信息。至于模型拟合的霍斯默和莱默苏检验，NOMREG 却用皮尔逊和偏 χ^2 统计，这是基于共变规则的数量，而非个案的数目。两者中，一般认为偏 χ^2 比 logistic 回归的皮尔逊 χ^2 能提供更多信息。偏 χ^2 统计显示较好的模型拟合，与图 3.1 的 Hosmer-Lemeshow 检验相似，分类表也一致。

图 3.2 和图 3.1 有一个较明显的不同，就是在图 3.2 参数估计(Parameter Estimates)的系数。在 LOGISTIC REGRESSION 中，我们要预测因变量的第二类别(第一类别是参考类别)，但我们可以选择任一类别作为参考类别，而在图 3.1 中，第一类别被选为参考类别，无论因变量和自变量，NOMREG 都默认最后一类为参考类别。因此，有些系数看起来就不一样，但其实它们与图 3.1 都是描述同样的事件。根据图 3.2 的结果，接触越多违法朋友的人越可能吸食大麻；违法观念越强烈，越倾向于不吸食；男性倾向于不吸食大麻(注意，这里是预测因变量的第一类别)。在图 3.2 中，参考类别是“白人”而不是“其他”，但两个 logistic 回归系数都没有显著性。

Case Processing Summary

		N
PMRJ5	0.00 no	147
	1.00 yes	80
SEX	1 MALE	110
	2 FEMALE	117
ETHN	1 white	175
	2 black	37
	3 other	15
Valid		227
Missing		30
Total		257

Parameter Estimates

PMRJ5		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	Upper Bound
.00 no	Intercept	0.978	2.122	0.212	1	0.645			
	EDF5	-0.407	0.069	34.347	1	0.000	0.666	0.581	0.763
	BELIEF4	0.118	0.060	3.903	1	0.048	1.125	1.001	1.265
	[SEX=1]	1.515	0.405	14.013	1	0.000	4.549	2.058	10.054
	[SEX=2]	0(a)	.	.	0	.	.	.	.
	[ETHN=1]	0.772	0.745	1.075	1	0.300	2.164	0.503	9.313
	[ETHN=2]	0.527	0.841	0.393	1	0.531	1.694	0.326	8.797
	[ETHN=3]	0(a)	.	.	0	.	.	.	.

a This parameter is set to zero because it is redundant.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	279.706			
Final	171.450	108.257	5	0.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	261.438	151	0.000
Deviance	158.266	151	0.306

Pseudo R-Square

Cox and Snell	0.379
Nagelkerke	0.522
McFadden	0.367

Classification

Observed	Predicted 0.00 no	1.00 yes	Percent Correct
0.00 no	134	13	91.2%
1.00 yes	28	52	65.0%
Overall Percentage	71.4%	28.6%	81.9%

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	171.450	0.000	0	.
EDF5	226.999	55.549	1	0.000
BELIEF4	175.535	4.085	1	0.043
SEX	187.461	16.011	1	0.000
ETHN	172.607	1.158	2	0.561

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

图 3.2 SPSS NOMREG logistic 回归的输出结果

图 3.3 显示部分 SAS 的输出结果，也是图 3.1 的模型。图 3.3 中 SAS 与图 3.1 中 SPSS 的输出结果有一点不同，就是处理 ETHN 的方法不一样。SAS 必须在分析时把 ETHN 转化成设计变量，因为 SAS 假定 PROC LOGISTIC 的自变量有一个真实值(SAS, 1989:1079)。而且 SAS 无法把 ETHN 当做一个变

```
Data Set: WORK.DATA1                          Response Profile
Response Variable: PMRJ5
Response Levels: 2                            Ordered
Number of Observations: 227                   Value    PMRJ5     Count
Link Function: Logit                              1        1         80
                                                  2        0        147

WARNING: 30 observation(s) were deleted due to missing values for the response or explanatory variables.
```

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	296.616	198.359	.
SC	300.041	218.909	.
-2 LOG L	294.616	186.359	108.257 with 5 DF (p=0.0001)
Score	.	.	89.457 with 5 DF (p=0.0001)

RSquare =0.379 Adjusted RSquare =0.522

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.7498	2.0285	0.7441	0.3883	.	.
EDF5	1	0.4068	0.0694	34.3468	0.0001	0.954476	1.502
BELIEF4	1	-0.1179	0.0597	3.9033	0.0482	-0.256713	0.889
SEX	1	-1.5148	0.4047	14.0130	0.0002	-0.418313	0.220
BLACK	1	0.2451	0.5080	0.2327	0.6295	0.050013	1.278
OTHER	1	0.7720	0.7446	1.0748	0.2999	0.105968	2.164

Association of Predicted Probabilities and Observed Responses

```
Concordant = 88.2%        Somers' D = 0.766
Discordant = 11.6%        Gamma     = 0.767
Tied       =  0.2%        Tau-a     = 0.351
(11760 pairs)             c         = 0.883
```

Hosmer and Lemeshow Goodness-of-Fit Test

Group	Total	GRP = 0 Observed	GRP = 0 Expected	GRP = 1 Observed	GRP = 1 Expected
1	24	24	23.31	0	0.69
2	24	21	22.58	3	1.42
3	25	24	22.38	1	2.62
4	24	22	20.34	2	3.66
5	23	18	18.10	5	4.90
6	23	13	16.13	10	6.87
7	23	14	13.08	9	9.92
8	23	7	7.68	16	15.32
9	23	3	3.13	20	19.87
10	15	1	0.28	14	14.72

Goodness-of-fit Statistic = 8.754 with 8 DF (p=0.363)

图 3.3 SAS PROC lOGISTIC 输出结果

量，必须把它分成几个设计变量，即检验分析不同的模型有没有 ETHN 这个自变量的模型。另外，参数估计、标准差、沃尔德统计、p 值、霍斯默和莱默苏检验、R^2_M（图 3.3 的 R 平方；SPSS 的考克斯—斯内尔）、R^2_N（图 3.3 的调整后的 R 平

方;SPSS 的 Nagelkerke)和－2 对数似然统计与 SPSS LOGISTIC REGRESSION 实际上是一致的。

解释非标准化 logistic 回归系数

图 1.4 显示二元 logistic 回归分析的结果。从图 1.4 中,我们得到方程式 logit(PMRJ5) = 0.407(EDF5) − 5.487。当 EDF5 等于最大值(29)时,方程式变成 logit(PMRJ5) = 0.407(29) − 5.487 = 6.316。若 EDF5 等于最小值(8)时,logit(PMRJ5) = 0.407(8) − 5.487 = −2.231。把 logit 转换成概率,当某人接触违法朋友的分数是 29 分,其使用大麻的概率等于 $e^{6.316}/(1+e^{6.316}) = 0.998$;当某人接触违法朋友的分数是 8 分,其使用大麻的概率变成 $e^{-2.231}/(1+e^{-2.231}) = 0.097$。当某人接触违法朋友的分数达到最高分,他/她一般也会使用大麻。当某人接触违法朋友的分数是最低分时,他/她使用大麻的相对概率小于 10%,虽可算是低概率事件,但比绝不可能使用大麻还差得多。若接触违法朋友的分数刚好是平均数,logit(PMRJ5) = 0.407(12) − 5.487 = − 0.603,使用大麻的概率是 $e^{-0.603}/(1+e^{-0.603}) = 0.354$,该数字接近这些 16 岁访问者使用大麻的非条件概率 ($P = 0.357$)。

正如线性回归系数,logistic 回归系数可理解成当自变量转变一个单位时,因变量 logit(Y)的变化。但 $P(Y = 1)$ 的变化与自变量的关系并不是线性的。曲线的斜率各不相同,它取决于自变量的数值。可以通过检测点与点间 $P(Y = 1)$ 的变化去计算不同点之间的斜率。例如,EDF5 由 8 到 9,概率由 0.097 增加至 0.101,为 0.004 的斜率,或 EDF5 由 28 增至 29,$P(Y = 1)$ 由 0.997 变成0.998,为 0.001 的斜率。EDF5 = 12 与 EDF5 = 13 之间的大麻使用概率由 0.354 变成 0.451,为 0.097 的斜率,这个改变比在最高或最低点时,每一个单位的改变大几倍。

logistic 回归系数的解释与含有数个自变量的模型差不多。图 3.1 大麻使用频率与自变量的关系的方程式为:logit(PMRJ5) = 0.407(EDF5) − 0.118(BELIEF4) − 1.514(SEX) + 0.245(BLACK) + 0.772(OTHER) − 1.749,其中变量 BLACK 和

OTHER 分别是指 ETHN(1)和 ETHN(2)。谈到单独系数，当 EDF5 每增加一个单位，logit(PMRJ5)增加 0.407。BELIEF4 每增加一个单位，logit(PMRJ5)减少 0.118。男性的 PMRJ5 的 logit 减少 1.514(记住，这个样本中男性比女性更少地使用大麻)。种族效应没有显著性[16]。

把特定个案的数值代入方程式就可以预测个案的情况，例如一个非裔美国女性 (BLACK = 1，OTHER = 0) 强烈认为违法是错误行为 (BELIEF4 = 25)，同时很少接触违法朋友 (EDF5 = 10)，则 logit(PMRJ5) = 0.407(10) − 0.118(25) − 0.1514(0) + 0.245(1) + 0.772(0) − 1.749 = −0.384。相应的大麻使用概率为 $e^{-0.384}/(1+e^{-0.384}) = 0.405$。另外，一个非西班牙欧裔美国男性 (BLACK = 0，OTHER = 0)，相信违法是错误的 (BELIEF4 = 20)，且其接触违法朋友(EDF5 = 15)的程度是中度，则 logit(PMRJ5) = 0.407(15) − 0.118(20) − 1.514(1) + 0.245(0) + 0.772(0) − 1.749 = 0.482，其相应的大麻使用概率 $e^{0.482}/(1+e^{0.482}) = 0.618$。

实质意义和标准系数

接触违法朋友增加一个单位的真正含义是什么？观念增加一个单位与一个单位的接触违法朋友真的等同吗？我们该不该认为观念每个单位的变化(原则上，范围由 7 至 28)等同于性别每个单位的变化(只有 0 或 1，范围只有 1)呢？这个问题在线性回归(因变量是大麻使用频率)或 logistic 回归(因变量是是否使用大麻)都会被问到。以不同的单位或量表量度自变量时，我们若想比较因变量与不同自变量的关系强度，通常会在线性回归中使用标准系数。同样的道理也适用于 logistic 回归，我们也可以考虑用回归分析的标准系数。

标准系数计算的单位是标准偏差。标准系数是指随着自变量每增加一个标准偏差，因变量的标准偏差会改变多少。在线性回归中，因变量 Y 与自变量 X 的标准系数 b^*_{YX} 可以从 Y 与 X 的非标准系数 b_{YX} 和他们的标准偏差 s_Y 和 s_X 计算出来：$b^*_{YX} = (b_{YX})(s_X)/(s_Y)$。另外，在回归分析之前，把 Y 和 X 标准化，即以 Y 和 X 减去它们的平均数再除以它们的标准偏差，$Z_Y = (Y-\overline{Y})/s_Y$ 和

$Z_X = (X - \overline{X})/s_X$ 得出 Y 和 X 之间的标准系数。

如果变量接近正态分布，99.9865%的个案都会落在 6 个标准偏差范围内（两边各 3 个标准偏差），而 99.999999713 的个案会落在 10 个标准偏差范围内。因此，自变量改变 1 个标准偏差意味着其可能值的范围会改变 1/8（小样本的 1/6，样本的 1/10）。根据 Chebycheff 不平等定理（Bohrnstedt & Knoke，1994：82—83），任何一个分布——即使是非常不正态的分布——至少 93.75%的个案会在平均数的 8 个标准偏差内，或者 96%的个案会在 10 个标准偏差内。因此直觉上，1 个标准偏差的变化足以影响其他变量（如果自变量对因变量有影响的话），但也不至于让原本微弱的关系变得明显，即使这一分布相当程度地偏离正态分布。如果我们使用通用单位（标准偏差或范围的 1/8）去量度所有自变量与因变量的关系，就算自变量本身测量的单位不一致，也可以比较他们对因变量的影响。

在 logistic 回归分析中，计算标准回归系数是复杂的，因为不是计算 Y 的数值，而是 Y 出现某个可能值或其他可能值的概率，这就是 logistic 回归方程所预测的。在 logistic 回归的因变量不是 Y，而是 logit(Y)，logit(0)的观测值是 $-\infty$，而 logit($+\infty$) $=+\infty$，也不能计算平均值和标准偏差。虽然我们不能直接计算观测值 logit(Y)的标准偏差，但我们利用 logit(Y)的预测值，间接计算其标准偏差，并可解释方差 R^2。第 2 章曾提过，$R^2 = \mathrm{SSR/SST}$。除以 N（样本就是 $N-1$），我们得到 $R^2 = (\mathrm{SSR/N})/(\mathrm{SST/N}) = s_{\hat{Y}}^2/s_Y^2$，也可写成 $s_Y^2 = s_{\hat{Y}}^2/R^2$。将 logit($Y$)代替 Y 和 logit($\hat{Y}$)代替 $\hat{Y}$，我们可根据 logit(Y)预测值的标准偏差和解释方差去计算 logit(Y)的方差。因为标准偏差是方差的平方根，我们可以估计标准 logistic 回归系数是：

$$b_{YX}^* = (b_{YX})(s_X)/\sqrt{s_{\mathrm{logit}(\hat{Y})}/R^2} = (b_{YX})(s_X)(R)/s_{\mathrm{logit}(\hat{Y})} \qquad [3.1]$$

b_{YX}^* 就是标准 logistic 回归系数，b_{YX} 是非标准 logistic 回归系数，s_X 是自变量 X 的标准偏差，$s_{\mathrm{logit}(\hat{Y})}^2$ 是 logit($\hat{Y}$)的方差，$s_{\mathrm{logit}(\hat{Y})}$ 是 logit($\hat{Y}$)的标准偏差，R^2 是判定系数。

用现时 SAS 和 SPSS 去计算标准 logistic 回归系数，需要以下这些必要的

步骤：

1. b：计算 logistic 回归模式以得出非标准 logistic 回归系数。储存 logistic 回归模型预测 Y 值。

2. R：利用 Y 的预测值去计算 R^2、R、η^2 或 η（这些量度都一样，用任何一个都可以）。

3. 利用 Y 的预测值去计算 logit（Y）的预测值，利用方程式 $\log(\hat{Y}) = \ln[\hat{Y}/(1-\hat{Y})]$。

4. $s_{\text{logit}(\hat{Y})}$：计算 logit($\hat{Y}$)的描述统计，包括标准偏差。

5. s_X：计算方程式中所有自变量的标准偏差，只包括模型中出现过的个案（换言之，当计算描述统计时，利用 listwise 删除缺失值）。

6. 把 b，R（或 η），s_X 和 $s_{\text{logit}(\hat{Y})}$ 代入方程式[3.1]算出 b^*。

标准 logistic 回归系数 $b^* = bs_X R/s_Y$ 的解释很直接，与线性回归的标准系数相似：当 X 增加 1 个标准偏差，会导致 logit(Y)改变 b^* 个标准偏差。在图 1.4 的模型中，EDF5 的标准偏差、logit($\hat{Y}$)的标准偏差和 η 都是分开计算的。EDF5 的标准偏差是 4.24，logit($\hat{Y}$)的标准偏差是 $s_{\text{logit}(\hat{Y})} = 1.72$，$R = \eta = 0.5871$，$b = 0.4068$。在方程式 3.1 中，$b^* = (0.4068)(4.24)(0.5871)/1.72 = 0.591$。换言之，EDF5 增加 1 个标准偏差，logit(PMRJ5)随之增加 0.591 个标准偏差。

表 3.1 概括了 SAS PROC LOGISTIC 和 SPSS LOGISTIC REGRESSION 的输出结果，还有已说明方差和预测效率的量度。

因变量与自变量的关系有显著性：$G_M = 108.257$，自由度 $= 5$，$p = 0.000$。因变量与自变量的相关强度 $R_L^2 = 0.367$ 和 $R^2 = \eta^2 = 0.435$（后者来自图 3.1），表示因变量与它的预测变量有中等强度的相关性。预测效率指针也显示模型的预测力不错：$\lambda_p = 0.488$，$\tau_p = 0.604$，两者都有显著性 $p = 0.000$。

比较表 3.1 和图 3.3 发现，表 3.1 的标准系数与图 3.3（SAS 分析）的标准系数估计不同。这是因为在 SASlogistic 回归系数的计算中：$b^*_{SAS} = (b)(s_X)/(\pi/\sqrt{3}) = (b)(s_X)/1.8138$，$\pi\sqrt{3}$ 是标准 logistic 分布的标准偏差（正如

表 3.1 是否使用大麻的 logistic 回归分析结果

因变量	相关/预测效率	自变量	非标准 logistic 回归系数 (b)	b 的标准误差	b 的显著性	标准 logistic 回归系数
PMRJ5	$G_M = 108.257$ ($p = 0.000$)	EDF5	0.407	0.069	0.000	0.531
	$R_L^2 = 0.367$	BELIEF4	−0.118	0.060	0.048	−0.143
	$R^2 = 0.435$	SEX(男)	−1.514	0.405	0.000	−0.233
	$\lambda_p = 0.488$	ETHN			0.552	
		黑人	0.245	0.508	0.630	0.028
		其他	0.772	0.745	0.300	0.059
	$\tau_p = 0.604$	截距	−1.749	2.028	0.388	/

标准正态分布的标准偏差 1)。SAS 只提供部分"标准",非完整的标准系数,没有考虑到 Y 或 logit(Y)的真实分布。另一种方法是只把自变量标准化。无论用 SAS 还是偏标准化的独立方法,与自变量对因变量的完全标准系数的排列是一样的,但很少会有人发现这个系数比完全标准系数 b^* 容易出现大于 1 或小于 −1 的情况,即使是在没有共线性(collinearity)或其他问题的情况下。完全标准的优点主要有:(a)相应的建构和说明与线性回归的标准系数相同;(b)可以把解释线性回归的标准系数的标准,同样应用于 logistic 回归。

如果我们简单地根据非标准化 logistic 回归系数(或相应地,根据发生比或概率),去评估自变量与 PMRJ5 的关系,性格 SEX 的影响最大,接着是 EDF5 和 BELIEF4,而 ETHN 则没有统计意义。但如果用标准系数,EDF5 就是最大(0.531,表 3.1),接下来是 SEX(−0.233)和 BELIEF4(−0.143)。换言之,(a)EDF5 增加 1 个标准偏差,logit(PMRJ5)相应增加 0.531 个标准偏差;(b)BELIEF4 增加 1 个标准偏差,logit(PMRJ5)相应增加 0.143 个标准偏差;(c)性别增加 1 个标准偏差("变得更男性化"),logit(PMRJ5)相应减少 0.418 个标准偏差。ETHN 在统计上不是 PMRJ 有意义的预测变量,它每转变 1 个标准偏差,logit(PMRJ5)就会相应改变 1/10 个标准偏差。

SEX 和 ETHN 增加一个标准偏差并不是指男女间或不同种族背景间的差异，这种差异在非标准 logistic 回归系数中反映出来。标准 logistic 回归系数的真正用途是通过把预测变量转化成同单位，比较其效果的强度。在描述实质结果时，像 EDF5 和 BELIEF4 这些没有量度单位的变量，选用标准系数比较合理。但像 ETHN 和 SEX(相应种族和性别间的差异)这种分类变量以及变量具有自然量度单位(寸、公升、元、事件数目)，非标准或指数化的系数比较合理。

指数化系数(exponentiated coefficients)或发生比数比

图 3.1 方程式内变量的最后一栏、图 3.2 中的参数估计(parameter estimates)的第三栏至最后一栏，以及图 3.3 最大似然估计分析的最后一栏，都是每个系数的发生比数比，如 SPSS 的 Exp(*B*)和 SAS 的发生比数比。当自变量增加 1 个单位，我们可将发生比数比乘以使用大麻的发生比(概率除以 1 减去此概率)，当发生比数比大于 1，表示当自变量增加时，使用大麻的发生比也会增加；当发生比数比小于 1，表示当自变量增加时，使用大麻的发生比也会减少。例如 EDF5 增加 1 个单位，成为大麻用户的发生比会增加 50.2%(成为大麻使用者的发生比乘以 1.502)。BELIEF4 增加 1 个单位，成为大麻用户的发生比会减少 11.2%(成为大麻使用者的发生比乘以0.889，比 1 少 0.112)。

这点很重要，发生比数比不是因变量和自变量关系的另类衡量。它所包含的信息与 logistic 回归系数或概率一样，只是表达形式有所不同。但发生比数比不能被当做标准 logistic 回归系数去比较各自变量对因变量影响的强度，如果所有发生比数比转换成大于 1(或小于 1)，就与非标准 logistic 回归系数排序相同，由最强排列到最弱。发生比数比跟 logistic 回归系数一样，只是形式不同，也没有更多的补充资料。

很多人错误地把发生比数比等同于风险率，有些人认为在某种程度的限制下，两者近乎相等(基数率少于 0.10)。一般而言，用发生比数比代表风险率会高估相关强度。男性的发生比数比 0.22(见图 3.1 和图 3.3)并不表示大麻使用

的风险比女性多 1/5，或女性的发生比数比 4.5(见图 3.2)也不表示女性使用大麻的风险是男性的 5 倍。要比较男女使用大麻的相对风险，必须利用模型去计算概率，假设已知另外一个数值。举例来说，白人男性和女性进行比较，EDF5 平均水平等于 12，BELIEF4 等于 27，女性的概率 $e^{0.407(12)-0.118(27)-1.749}/(1+e^{0.407(12)-0.118(27)-1.749})=0.487$，男性的概率 $e^{0.407(12)-0.118(27)-1.514-1.749}/(1+e^{0.407(12)-0.118(27)-1.514-1.749})=0.173$。男女的相对风险约为 3 倍(女∶男=2.8∶1 或男∶女=0.35∶1)，而不是 4 至 5 倍。

分类预测变量：对比和解释

图 3.1 和表 3.1 用 0 和 1 去代表 ETHN 的不同可能值，这叫做指针编码(indicator coding)，因为它显示类别特征存在或不存在。指针编码是处理 logistic 回归分析的设计变量的方法之一。另一个方法是 SPSS LOGISTIC REGRESSION 的简单编码，logistic 回归系数中设计变量的简单对比与指针编码一致，只是截距有所不同。

另一个方法是离差编码(deviation)，SPSS LOGISTICS REGRESSION 默认的选项。离差编码将每个设计变量与自变量整体效果比较，类似比较回归或方差分析中三个分类的平均。在 logistic 回归中，离差编码量度每组 logit 与整体样本的平均 logit 存在偏差，因此，参考组不再是数字 0。反之，它的系数等于其他类别的系数总和的负数。如果计算机时间比人工要贵，那人工计算被删除的系数会比较合理。但如果用个人计算机或免费计算机，把不同参考类别的两个模型都算出来就比较合适，这样不但可以得出全部三个分类的系数估计，还可得到删除的分类的标准偏差和统计显著性[17]。ETHN 的离差编码模型的综合概要列在表 3.2 中。

除了改变 ETHN 个别系数，用不同指针变量的编码方案不会影响分析结果。种族组别的排列与图 3.1 一样(非西班牙欧裔美国人使用大麻概率最低，其他的概率较高)，而不像表 3.1，表 3.2 补充了非裔美国人，如果种族效果呈显著性，那么，整体上与大麻使用概率成负相关。表 3.1 重组表 3.2 的指标对

比系数，把表 3.2 中的每个系数减去非西班牙欧裔美国人系数－0.3388。因此，离差编码所提供数据与指针编码相似，但参考点不是一个类别，而是平均效果。

表 3.2 大麻使用普遍程度的 logistic 回归：种族的偏差编号

因变量	相关/预测效率	自变量	非标准 logistic 回归系数 (b)	b 的标准误差	b 的显著性	标准 logistic 回归系数
PMRJ5	$G_M = 108.257$ ($p = 0.000$)	EDF5	0.407	0.069	0.000	0.531
	$R_L^2 = 0.367$	BELIEF4	−0.118	0.060	0.048	−0.143
	$R^2 = 0.435$	SEX(男)	−1.514	0.405	0.000	−0.233
	$\lambda_p = 0.488$	ETHN			0.552	
		白人	−0.339	0.319	0.289	−0.044
		黑人			0.810	−0.011
		其他			0.388	0.033
	$\tau_p = 0.604$	截距	−1.140	2.042	0.490	/

SPSS logistic 回归分析中的其他对比包括赫尔默特(Helmert)、反向赫尔默特(reversed Helmert)、多项(polynomial)、重复和特别对比。赫尔默特、反向赫尔默特、直角正交(orthogonal)和反复对比适用于检验定序预测变量不同类别的效果与类别排序是否一致。多项对比用于检验线性与非线性效果。用于定序变量的不同对比法不会影响模型拟合或分类、定序变量的显著性。这一结果显示，当类别间有自然断裂或明显线性或单向性时，变量编码可做适当的转变。最简单的定序对比是反复对比(repeated contrast)(SPSS)和形象对比(profile contrast)(SAS)，这些方法是把自变量的每个分类，除了第一个(参考类)之外，与之前一个类别做比较。通过检查类别的系数，就可以知道自变量和因变量是否存在单向或线性的关系。如果正和负系数出现非系统规则，表示有非线性、非单向的关系存在，自变量最好视为名义变量而非定序变量。

当设计变量用来表示单一名义变量的效果时，设计变量就要被视为一个组群，而不是单独变量。个别设计变量的显著性应只视为一个组群对因变量的显著性。个别设计变量的统计意义应可解释为对是否与参考类别（指针编码）或分类变量的平均效应（离差编码）有显著性的差异，假定分类变量一开始就有显著效应。SPSS 检验名义变量效果的显著性（所有设计变量都包括）。SAS 也有类似的检验，使用逐步程序去比较有和没有包括所有代表种族的指标变量的名义变量模型。而定序对比设计变量的整体显著性只表示当分类变量被视为名义变量时，是否对因变量有显著影响；个别系数的显著性可能提供有关分类预测变量和因变量之间关系的重要信息，即使分类变量对因变量没有呈现显著影响。

交互作用

在一些统计软件中，我们只需加入指定交互作用的变项，软件就会把它放在方程式内计算，并得出其与因变量的相关强度及显著性。另一些软件就必须分开计算交互作用的变项，才可加入模型中。唯一的复杂性是当涉及与名义变量（有两个或以上分类）的交互时，就需要比较是否包括所有交互项的模型，以确定是否具有统计和实质性的意义。在线性回归中，一个保守估计交互作用的统计意义方法，是看在模型中加入交互变项引致 R^2 有没有显著的变化，R^2 变化的幅度可以反映实质意义（加入交互变项能增加多少预测能力）。在logistic 回归中，相应的准则也是看 G_M 和 R^2_L 变化的显著性。

表 3.3 显示表 3.2 模型加入了两个交互变项：SEX 和 EDF5 的交互变项（检验男女接触违法朋友的影响是否不同）、SEX 和 BELIEF4 的交互变项（检验男女之违法观念的影响是否不同）。这是检验曲线的部分斜率的差异，曲线代表男女 PMRJ5 与 EDF5 和 BELIEF4 的关系，两个交互变项都没有显著性。但这两个变项加起来，G_M 有边缘性意义（4.656，$p = 0.098$），R^2_L 有轻微的增长（0.016），λ_p 减少（−0.025），τ_p 也减少（−0.019）。

表 3.3 性别与违法观念和接触违法朋友的交互作用检测

因变量	相关/预测效率	自变量	非标准 logistic 回归系数 (b)	b 的标准误差	b 的显著性	标准 logistic 回归系数
PMRJ5	$G_M = 112.913$ ($p = 0.000$)	EDF5	0.549	0.126	0.000	0.662
	$R_L^2 = 0.383$	BELIEF4	−0.161	0.088	0.067	−0.180
	(R_L^2 的变化 $= 0.016$)	SEX(男)	−1.516	0.408	0.000	−0.215
	$R^2 = 0.435$	ETHN			0.459	
	(没有变化)	黑人	−0.303	0.516	0.558	−0.032
		其他	0.891	0.761	0.242	0.063
	$\lambda_p = 0.463$ (变化 $= -0.025$)	SEX × Z_{EDF}	−0.919	0.638	0.150	−0.187
	$\tau_p = 0.585$	SEX × Z_{BELIEF}	0.451	0.491	0.358	0.084
	(变化 $= -0.019$)	截距	−2.132	3.122	0.495	/

注：为避免出现共线性，交互项的 EDF5 和 BELIEF4 已标准化。因为当加入交互变项时，预测表多了两个假阴性(false negative)。最合理的结论是违法观念和接触违法朋友对男女是否使用大麻的影响是一样的。

逐步 logistic 回归

逐步回归原则上可用于表 3.3 的交互作用分析，以检验使用大麻普及性的模型中是否适合加入交互变项。线性或 logistic 回归模型通常包括或删除某预测变量是根据计算器的算式程序，而不是直接由研究人士来决定的。因此，有些人不认同逐步方法是用于探索性研究的好方法(Agresti & Finlay, 1997: 527—534; Hosmer & Lemeshow, 198:106)；其他人批评它一开始就忽略要研究的某些现象(Studenmund & Cassidy, 1987)。我们在此不深入讨论逐步程序的争论，认为使用计算机控制选择变量的逐步程序不适用于理论检验，因为它利用数据的随机变化产生出来的结果往往是千变万化且难以在不同样本中重

复得出相同的结果。

支持者则认为在两种情况下逐步程序是很有用的:纯粹预测和探索性研究。在纯粹预测研究中,由于不关注因果关系,只要确定含有一组预测变量的模型,就可以准确地预测某些现象。例如,大学招生办公室可能想知道哪些因素能精确地预测到高校成功与否,不是用于做理论的发展,而纯粹是为选择生源。探索性研究有可能关注到预测和解释一个现象的理论建构和发展,当这一现象是新现象时,又很少研究文献或"理论"。例如,沃福德(Wofford)、埃利奥特(Elliott)和梅纳德(Menard, 1994)使用逐步程序的探索性研究。

沃福德等人研究家庭暴力的连续性,样本是从全国家庭中抽样 18 岁至 27 岁的青年男女。根据家庭暴力的文献,他们的研究分析包括了 26 个预测变量。作为研究的一部分,在 1984 年访问时,受访者报称,受害者或肇事者会在 1987 年接受重访,看看家庭暴力是否继续或自 1984 年就停止了。总共有 108 名妇女(原样本 807)在 1983 年接受访问时称自己是家庭暴力的受害者,1986 年又接受重访。沃福德等人建构了一个 logistic 回归模型,包括所有 26 个预测变量。因为在这方面的理论并不发达,加上个案数目相对于解释变量的数目较少,所以他们采用了逐步回归分析法。

逐步回归选用了向后淘汰法(backward elimination),而不是向前包括法(forward inclusion)。在某些情况下,变量可能呈显著性,因为另一个变量被控制或保持常数,这叫做抑制效应(suppressor)(Agresti & Finlay, 1997:368)。向前包括法作为逐步回归的一个方法的缺点就是,抑制效应可能排除某些变量。但向后淘汰法都已包含两个变量,所以未能找到存在的关系的风险较低。线性回归的向后淘汰法和向前包括法会产生同样的结果,但当向后淘汰法发现向前包括法找不到的关系时,两个结果就会不同。

为进一步防止未能找到一个存在的关系,应放松通用的 0.05 统计意义标准。阿菲菲(Afifi)等人(1977)的向前逐步回归研究认为 0.05 太低,常常不包括一些重要的变量。他们建议将统计意义标准的范围设在 0.15 至 0.20。当原假设是真的,拒绝原假设的风险会增加(找到一个确实不存在的关系),但当原假

设是假的，没有拒绝原假设的风险会降低（找不到确实存在的关系）。在探索性研究中，与理论检验恰恰相反，它往往比淘汰劣质预测变量更注重寻找良好的预测变量。沃福德等人审查了三种模型：一个含全部变量的全 logistic 回归方程模型；一个排除 p 大于 0.10 的变量的简化模型（实际上，这与使用 0.15 或 0.20 的标准一样）；以及一个排除 p 大于 0.05 的变量的简化模型。结果列于表 3.4。

表 3.4 持续的婚姻暴力受害者（女性）

$N=108$	模型 1 包括所有变量	模型 2 最大 $p=0.100$	模型 3 最大 $p=0.050$
模型 $\chi^2 G_M$ （自由度）	30.254 (28df)	21.284 (7df)	4.472 (1df)
G_M 的显著性	0.351	0.003	0.034
D_M	119.429	128.298	145.211
G_M 变化 （与前模型相比的自由度变化）	—	8.870 (21df)	16.812 (6df)
G_M 变化的显著性	—	0.99	0.010
R_L^2	0.202	0.150	0.030
τ_p	0.481	0.509	0.145

模型 2 中的预测变量

自变量	b	标准误差	p（基于似然率统计）
接受社会福利	1.88	0.95	0.03
社会阶层背景	−0.03	0.02	0.05
以前受到轻微的侵犯	1.24	0.53	0.02
以前受到严重的侵犯	−1.07	0.62	0.08
目睹父母的家庭暴力	−1.70	0.64	0.00
被另一半严重暴力对待的频率	0.12	0.06	0.05
寻求专业帮助	0.88	0.53	0.09

表 3.4 第一部分比较了三种模型。全模型的 G_M 没有显著性，表示预测变量在解释因变量的贡献方面不比机会率好。模型的 χ^2 未达至统计意义的部分

原因是样本太小，另一种可能是大量变量包括在模型中。模型 2($p < 0.10$) 的 G_M 很小但有显著性，模型 3 也是如此 ($p < 0.05$)，但他们都只有一个预测变量。模型 χ^2 的变化(或等同的，D_M 的变化)，从模型 1 至模型 2 和从模型 1 至模型 3 的变化都没有显著性。但是，模型 2 至模型 3 的变化在 0.01 水平上有显著性。全模型 1 的 R_L^2 是 0.20，模型 2 的 R_L^2 减少到 0.15，模型 3 的 R_L^2 只有 0.03。模型 1 的 τ_p 是 0.48；模型 2 增至 0.51，但模型 3 只有 0.14。

选用模型 2 做进一步的分析，是因为(1)降低模型的 G_M 有统计学意义，而完整模型没有；(2)模型 2 比模型 3 的模型拟合有更好的统计学意义，同时并不比全模型差；(3)模型 1 比模型 2 的 R_L^2 和 τ_p 的变化相对较小(相反的方向)，但模型 3 比模型 2 少得多。模型 2 的结果列在表 3.4 的下半部分。

表 3.4 显示了逐步 logistic 回归方法的几个要点，其中最重要的一点是，结果是非常初步的、尚未定论的。这是在寻找可能的预测变量，而不是具备说服力的理论检验。第二，逐步过程的重点在于比较全模型与降低模型。正如阿菲菲等人(1977)的建议，比较两者的拟合优度和预测效率统计，0.05 显著水平的标准似乎过于严格，需要选用一个更合理、更宽松的显著水平。第三，模型 2 所确认的预测变量是预测家庭暴力的颇好因素，但有些很容易受其他原因的影响(例如寻求专业帮助)。这些结果将促进进一步的理论发展和检验，但需要重复研究才能解释(最好有更明确的依据)和预测持续性家庭暴力的原因。

第 4 章 | 诊断 logistic 回归的介绍

当不符合 logistic 回归分析的假设时，计算 logistic 回归模型的效应可能会出现系数偏差、效率太低或无效的统计推断。偏差是指估计 logistic 回归系数出现过高或过低的系统性趋势，比起真正的系数来说太远或太近。低效率指系数有较大的标准偏差，相对于其数值来说，即使原假设是错误的，也将难以拒绝原假设(假设因变量和自变量不存在相关)。无效统计推断指计算 logistic 回归系数的显著性不准确。此外，高杠杆个案、样本的自变量有异常高或低的数值(不是因变量，它只有两个值)，或在指定的自变测量值上，因变量有离异值(outlier)，这都可能导致估计参数不相称。本章的重点是讨论不符合 logistic 回归假设所带来的后果、检测和修正方法，并简介审议检测异常值、高杠杆个案、logistic 回归具影响力的个案和处理方法。

设定误差

线性和 logistic 回归分析首要的假设是拟定的模型是正确的。

正确设定有两个元素：模型的函数形式是正确的，以及模型包含所有有关的自变量，没有无关的自变量。错误的模型指定可能得出有偏差的 logistic 回归系数，系数会出现系统性的高估或低估。在第 1 章中我们知道，应用线性回归分析于二分因变量会出现错误的模型指定，因此有必要研究 logistic 回归模型。logistic 回归模型的因变量是 logit(Y)与自变量呈线性组合，模型函数形式是不正确的。首先，logit(Y)可以与自变量呈非线性组合；第二，自变量之间的关系可能是相乘或交互，而不是相加。

logistic 函数的指定误差，相对于不同 S 型函数，并不算是个问题。奥尔德里发生比和纳尔逊(1984)指出，logit 模型(基于 logistic 功能)和概率模型(基于正态分布)的结果极为相似。霍斯默和莱默苏(1989:168)指出，logistic 回归模型是高度灵活的，概率介于 0.2 至 0.8，能产生与其他模型非常类似的结果。

删除相关变量和包括无关变量

包括一个或多个不相关的变量会增加参数估计的标准误差，即减少估计的效率，但系数不会有偏差。标准误差的膨胀幅度取决于模型中的不相关变量与其他相关变量的大小。如果与模型中的其他变量完全不相关的话，标准差不会增加，但实际上这种可能性极小。

删除相关变量会导致自变量系数的偏差，其程度要根据删掉的变量与 logistic 回归方程中的自变量的相关程度来确定。和线性回归(Berry & Feldman, 1985)类似，偏差的方向取决于被排除变量的参数与因变量效应的方向及包括在内的变量的方向。而偏差的幅度取决于被排除的变量和其他变量之间关系的强度。如果被排除变量的偏差与包括在内的变量不相关的话，系数不会偏差，但在实际情况中，这不可能发生。一般来说，偏差比低效率的问题更严重，但少量的偏差比严重的低效率好。

出现排除相关变量的错误，可能是因为现有的理论无法找出因变量的所有相关预测变量，或者理论上相关的变量都被删除了。如果错误地设定模型函数形式，就会出现删除变量的结果。非线性模型被断定成线性，计算上等同于排除了一个变量，一个代表因变量与自变量的非线性组合。非相加模型被断定为相加模型，相当于除去一个变量，尤其是建构模型中两变量的交互作用。无论被排除的变量是非线性的或是交互的，只有理论(或低到令人失望的 R_L^2 值)能帮助检测和纠正这个问题。当被排除变量真的是非线性或交互的，变量函数会出现在方程中，那检测和纠正问题可能会比较容易。

logit 的非线性

在线性回归模型中，因变量的变化与自变量每个单位的改变的关系是成常

数的，即自变量的回归系数。如果 Y 对 X 每改变一个单位的变化是依赖于 X 值(正如 Y 是二分变量)，那么这种关系是非线性的。相应的，当 logit(Y)为因变量，如果 logit(Y)的变量随着 X 每个单位的变化呈常数，同时不依赖于 X 值，我们说这个 logistic 回归模型是线性形式或者 logit 呈线性关系，以及 logit(Y)的改变与 X 每单位的变化等于 logistic 回归系数。如果关系在 logit 中不是线性的，logit(Y)的改变与 X 每个单位的变化就不是常数，而且依赖于 X 值。

检测因变量 logit(Y)与每个自变量之间非线性关系的几个可能的方法(Hosmer & Lemeshow, 1989:88—91)有：一是把每个自变量作为分类变量，并使用正交多项式的对比来检验二元 logistic 回归或多元 logistic 回归模型中的线性、二次、三次和高阶效应。如果自变量有大量的分类(例如 20 类)，标准误差倾向就会比较大，无论是线性或非线性效应可能都没有显著性，即使事实上有显著的线性效应存在。二是用霍斯默和莱默苏(1989:90)所描述的博克斯—蒂德韦尔(Box-Tidwell)转化，就是在方程式中加入 X ln(X)(X 乘以 X 的自然对数)。如果这个变量的系数是有显著意义的，那么证据显示 logit(Y)和 X 有非线性关系。霍斯默和莱默苏指出这个程序不能很敏锐地检验出弱小的偏脱机性，也不能指出确切的非线性形式。如果关系是非线性的，有必要做进一步调查以确定非线性的形式。

第三个方法也是霍斯默和莱默苏所建议的，即根据自变量的数值把个案分成不同组别，计算每组因变量 Y 的平均数，然后把每组 Y 的平均数转成 logit，再对自变测量值做图。对于自变量 X 的每一个数值 i，Y 的平均数是 $P(Y=1 \mid X=i)$ 的概率。但这个程序会出现一个问题，对任何 X 值来说，Y 是 1 或 0。如果是这样的话，我们无法计算 logit(Y)，它等于正负无限大。解决方法是把相邻类似但相等概率的类别组合起来。但是这可能会掩盖某些非线性的关系。另一个可能性是给那些平均数为 1 的组别指定一个任意大的平均数(例如 0.99)，给那些平均数为 0 的组别指定一个任意小的平均值(0.01)。这种方法有助于确定非线性的模式，而且图像显示有助于识别自变量或因变量与自变量组合的异常个案。

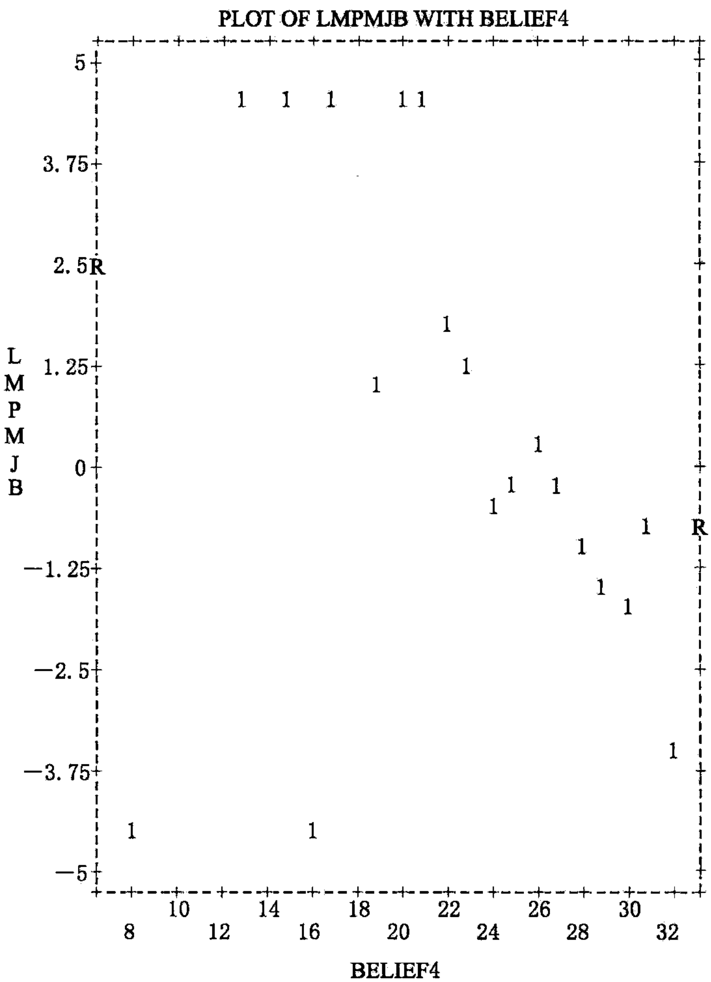
表 4.1 列出了非线性博克斯—蒂德韦尔检验的结果。表 4.1 的第一部分将这

两个非线性项 BTEDF = (EDF5)ln(EDF5) 和 BTBEL = (BELIEF4)ln(BELIEF4) 加到模型中。这两个非线性交互项都呈显著性（G_M 的变化 = 15.066，自由度 = 2，p = 0.005），但基于 G_X（每个非线性项的似然率统计），只有 BTBE 呈显著性(G_X = 9.932，自由度 = 1，p = 0.002)，BTEXP 的 G_X = 2.226，自由度 = 1，p = 0.136。当模型不包括 BTEXP 时，BTBEL 的系数仍有显著性(G_X = 12.839，自由度 = 1，p = 0.000)，包括 BTBEL，R_L^2 值增加 0.034(3.4%)。

表 4.1 博克斯—蒂德韦尔正态检验

因变量	相关/预测效率	自变量	非标准 logistic 回归系数 (b)	b 的标准误差	b 的显著性(基于似然率 G_X)	标准 logistic 回归系数
PMRJ5	G_M = 123.322 (p = 0.000)	EDF5	2.415	1.191	0.076	3.073
n = 227	R_L^2 = 0.418	BELIEF4	3.620	1.421	0.002	−4.272
	G_M 的变化 =	SEX(男)	−1.660	0.428	0.000	−0.249
	15.065(p = 0.001)	ETHN			0.516	
	(对基本模型)	黑人	0.333	0.536	0.535	0.036
	(表 3.1)	其他	0.859	0.817	0.293	0.063
	R_L^2 的变化 = 0.051	BTEDF	−0.551	0.325	0.136	−1.620
	(对基本模型)	BTBEL	−0.891	0.340	0.002	−4.172
		截距	−29.210	8.733	0.001	—
PMRJ5	G_M = 125.195 (p = 0.000)	EDF5	2.411	1.184	0.076	3.012
n = 225	R_L^2 = 0.427	BELIEF4	1.380	2.812	0.644	−1.503
	G_M 的变化 =	SEX(男)	−1.676	0.436	0.000	−0.249
	2.561(p=0.278)	ETHN			0.530	
	(对基本模型)	黑人	0.342	0.534	0.522	0.036
	(n = 225)	其他	0.823	0.814	0.312	0.060
	R_L^2 的变化 = 0.009	BTEDF	−0.555	0.323	0.137	−2.551
	(对基本模型)	BTBEL	−0.371	0.660	0.599	−1.716
	(n = 0.225)					
		截距	−15.541	17.134	0.364	—

图 4.1 显示了为什么 PMRJ5 与 BELIEF4 的关系是非线性的。因为 PMRJ5 的平均数用于计算每个 BELIEF4 的数值，而且 BELIEF4 几个值的 PMRJ5 平均数是 0 或 1，数值 1 被重新编码为 0.99，0 被重新编码为 0.01。接着，把每个平均数的 logit 对 BELIEF4 做出来。在图像的左下象限有两个离群值，都是单一个案，表示受访者有非常薄弱的违法观念，但报告没有使用大麻。除了这两个个案，图像不再出现大幅度偏脱机性的情况。表 4.1 第二部分证实了，删除



注：单独的分类自变量和截距都没有似然率统计，其显著性的计算是基于沃尔德统计，LMPMJB = 每个 BELIEF4 数值的大麻使用率平均数的对数，BELIEF4 = 相信违法是错误的。

图 4.1 logistic 回归的诊断：非线性检验

这两个个案后的分析显示，BTEDF 和 BTBEL 无论是分开还是组合在一起，都没有对 PMRJ5 的 logit 有显著的影响。这些个案是否应予删除或保留将留待后文讨论。

非相加性

如果因变量的变化是随着自变量的数值而改变的话，因变量与自变量的关系就是非线性。因变量的改变随自变量每单位的改变是取决于另一个自变测量值。例如，接触违法朋友这一变量每改变一个单位，薄弱至中等的违法观念的人（更容易受到同伴的影响）比起强烈违法观念的人（可能较不容易受朋辈的影响）在大麻使用频率或普及性上，可能会有更大的变化。相加性的检测并非如线性或 logistic 回归的非线性检测一样简单。除非理论提供了一些指引，我们通常会检验直观上有交互作用的变项，或所有可能的交互变项。后者对简单的模型比较可行，但当模型中的变项测量数目增加，就会变得很繁杂并导致越来越多的随机抽样误差出现的风险。

表 3.3 的例子检验一个交互效应，结果显示：男性与女性接触违法朋友和违法观念对其使用大麻普及性没有显著区别。这是一个连续预测变量（实际上是两个：EDF5 和 BELIEF4）与一个二分变量的例子。其他可能性包括两个分类变量（SEX 和 ETHN），或两个连续变量（EDF5 和 BELIEF4）的交互作用。

共线性

当自变量彼此相关会出现共线性（或共线性或多重）。完美共线性是指一个自变量与其他自变量完美线性相关。如果我们把每个自变量当做模型中的因变量，用其他自变量去预测，完美共线性会导致每个自变量得出 $R^2 = 1$。当完美共线性存在的话，就不可能取得唯一的回归系数估计，任何一个（无限可能的组合）都适用于线性回归系数或 logistic 回归系数。完美共线性是罕见的，除非被疏漏掉：包括三个变量，其中一个是其他两个的总和，便是一个例子。

但不完美的共线性是相当普遍的。自变量之间的任何相关都是共线性。

自变量中的共线性增加，线性和 logistic 回归系数也不偏差，而且效率一样（已知自变量之间的关系），但线性或 logistic 回归系数的标准误差会比较大。不可能有更多有效率而不偏差的估计，效率水平的估计可能会很差。低度共线性一般不存在问题，但高度共线性（至少有一个自变量的 $R^2 = 0.80$ 或以上）可能会造成问题，很高共线（至少一个自变量 $R^2 = 0.90$ 或以上）几乎肯定系数不显著，即使可能系数相当大。共线往往令线性和 logistic 回归的系数出现不合理的高：我们有一个粗略的准则，当标准 logistic 或线性回归系数大于 1 或非标准 logistic 回归系数大于 2 时，需确定是否有共线性存在。

线性回归的共线性检测非常简单直接。被广泛使用的软件程序的最标准回归程序提供每个自变量的 R^2 的最好或部分功能，当把该自变量视为因变量时，所有其他自变量可以被当做预测变量。例如，SAS 的 PROC 回归和 SPSS 回归的宽容统计 $1-R_X^2$，R_X^2 是每个自变量 X 用所有其他自变量去预测得到的决定系数。依据前文的简单准则，当宽容统计少于 0.20 时，要注意，当宽容少于 0.10，几乎可以肯定有严重共线性的问题。虽然 SAS 的 PROC LOGISTIC 或 SPSS LOGISTIC 回归没有宽容统计，但是可以通过线性回归模型计算出来，这一模型的因变量和自变量与 logistic 回归模型相同。因为关注自变量间的关系，因变量的模型函数形式与共线估计无关。

表 4.1 是两个非线性项 BTEDF 和 BTBEL 的模型，logistic 回归系数有点高，EDF5，BELIEF4，BTEDF 和 BTBEL 的标准 logistic 回归系数均大于 1。这表明非线性模型的共线性可能存在。表 4.2 列出共线性统计量估算方法是 OLS 回归。第一，称为“基本模式”是 logistic 回归模型（表 3.1 的模型）。第二，表 4.1 的上半部分包括非线性项 BTEDF 和 BTBEL 的“非线性模型”。对于基本模型，在这两个模型中，黑人和其他的 ETHN 设计变量的所有宽容统计都超过 0.70，表示没有严重的共线性。对于非线性模型，表 4.2 证明表 4.1 的标准化系数：性别和两个 ETHN 设计变量的宽容统计仍然很高，但 BTEDF 和 BTBEL 与 EDF5 和 BELIEF4 的宽容统计少于 0.01，表示有严重共线。共线性很容易被检测，但很少有有效的补救措施。删除涉及共线的变量会有遗漏变量的偏误，把共线性的变量组合成单一变量，例如，利用因素分析，理论（如有）用于

建构你的模型或数据收集的测量过程会出现错误,可以参考分析结果做进一步推论。岭回归(ridge regression)(Schaefer, 1986)允许我们通过增加变量的估计方差,产生有些偏差但实际效率高的估计,从而减少被解释的方差比例。或许最安全的策略是集中模型中所有变量的综合影响,同时承认高共线性的存在,个别预测变量的结论会不稳定。如想更详细地讨论共线性的补救办法,请看贝里和费尔德曼(1985:46—50)或福克斯(1991)的著作。简单地说,高共线性没有真正令人满意的解决办法。

表 4.2 共线检验

因变量	自变量	宽容度	
		基本模型	非线性模型
PMRJ5	EDF5	0.717	0.00249
	BELIEF4	0.707	0.00148
	性别(男性)	0.994	0.994
	ETHN		
	黑人	0.959	0.958
	其他	0.983	0.974
	BTEDF	—	0.00253
	BTBEL	—	0.00147

数值问题:零格数和完全分离

共线性的存在不一定表示模型或模型理论有什么不妥当。相反,问题在于研究数据(自变量之间具有高度相关性)。有两个问题与此相关:格数内的数值为零和完全分离。当因变量对自变量的一个或多个值都没有变异的话,格数数值会出现零。例如种族中的其他类别,如果所有受访者都报称使用大麻(或者没有使用大麻),那么大麻使用和种族关系列联表的格数会出现零。白人和黑人受访者的使用大麻者的发生比是 $1/(1-1)=1-0=+\infty$,同时 $\text{logit}=\ln(\text{odds})$ 也会呈 $+\infty$,无限大。如果这组使用大麻的普及率等于 0,发生比等

于 0/(1－0) ＝ 0，同时 logit ＝ ln(0) ＝－∞，无限小。当个案的发生比是 0 或 1 时，这不是一个问题，但根据分类自变测量值定义而得出整个组别或全部个案是 0 或 1 的话，该类别的系数就会出现非常高的估计标准误差（包括参考类别）。

分类变量，特别是名义变量会出现格数值为零的问题。对于连续变量和定序分类变量，通常自变量的 0 或 1 数值都有其代表的意义，原因是假定因变量和连续预测变量具有某种模式的关系（线性回归的线性，logistic 回归的对数），并可使用该模式去填补因变量对自变量的分布上的空白。我们无法假设分类变量有这种情况。相反，当我们发现分类预测变量的格数值为零时，必须选择：(1)接受高的标准误差和不确定的 logistic 回归系数；(2)重新编码自变量成更有意义的分类（可以将分类组合在一起或删除有问题的分类），以消除格数值为零的问题；(3)列联表的每格都加上一个常量以消除零数值。

和每个预测变量效应相比，如果我们更关注所有预测变量与因变量的整体关系，那么第一种选择就更容易接受。整体模型是否适合应不受格数值 0 的影响。第三种选择没有严重的弊病，但霍斯默和莱默苏(1989:127)提出，它可能不适合较复杂的分析。第二个选择会导致自变量的测量比较粗略，令其与因变量的关系强度出现偏差和趋向零。但是，如果有些分类自变量之间有概念的联系，或如果因变量对自变量的某些类别的分布都很类似，这可能是一个合理的选择，通常可被用在单元变量和二元变量的数据筛选中。本章一直有个隐藏的例子：种族编码。在最初的调查中，种族分为六大类：非西班牙欧裔美国人、非裔美国人、西班牙裔美国人、印第安人、亚裔美国人和其他。因为最后四类个案少，所以合成"其他"类别。如果保留原来的形式，格数值零的问题会严重破坏统计分析。

如果自变量太成功地预测因变量，就有完全分离的问题。logistic 回归系数及标准误差往往趋向为极大。因变量将被完美地预测到：$G_M = D_0$，$D_M = 0$，$R_L^2 = 1$。如果分开不是完全的（有时称为假完全分离），logistic 回归系数及其标准误差仍是非常大的。图 4.2 是个基于随意加上数字的假分离例子。

二元 logistic 关系会出现完全分离的现象，logistic 回归模型不能计算出

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13.003	0.654	0.872

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	42.448	1	0.000
	Block	42.448	1	0.000
	Model	42.448	1	0.000

Classification Table(a)

	Observed		Predicted TRUE		Percentage Correct
			0	1	
Step 1	TRUE	0	19	1	95.0
		1	1	19	95.0
	Overall Percentage				95.0

a The cut value is 0.500

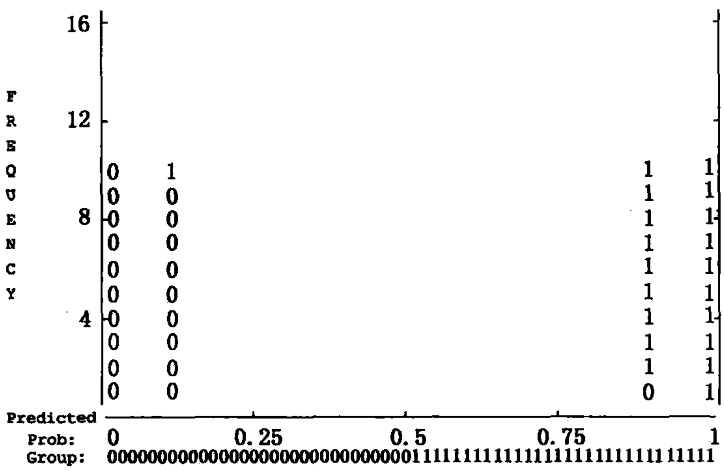
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P3	219.720	74.535	8.690	1	0.003	2.65E+95
	Constant	-109.860	37.275	8.687	1	0.003	0.000

a Variable(s) entered on step 1: P3.

Step number: 1

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for 1
The Cut Value is 0.50
Symbols: 0 - 0
1 - 1
Each Symbol Represents 1 Case.

图 4.2 近似完全分离

来。虽然完全分离本质上没有什么错(但毕竟我们努力实现达至完美的预测),但是实际上它往往会引起我们的怀疑,因为它在现实世界中几乎不会存

在。完全或假完全分离可能表示数据或分析出现问题，例如，变量与个案的数目太相近。

共线性、格数值为 0 和完全分离有相同的症状，就是标准误差非常大，往往但不一定出现大的系数（Hosmer & Lemeshow，1989）。因此会导致模型出现低效率的参数估计，但不知道导致偏差的参数或不准确的（与低效率相反）的推论的原因。格数值为零可以在 logistic 回归分析前，通过单项因素和二元变量分析查出来。完全分离可能表明没有错误需要纠正，或者就是辉煌的理论和分析上的突破，但最有可能存在问题。共线性是三个问题中最令人烦恼的，因为它表明不是理论或运作上有缺陷，就是数据出现问题，这些都扰乱了理论的检验。这一理论关注个别预测变量的效应，而不是所有变量的综合影响。像格数值为零，可以在 logistic 回归分析前检测到共线性（在良好多元回归软件的帮助下）。如果能查到问题，那么如何解决也是一门艺术而非科学。

残差分析[18]

在线性回归中，残差通常会写成 e 和 $e_j = Y_j - \hat{Y}_j$，就是个案 j 的 Y 观测和预测值的差别。这应该与预测误差ϵ_j区分开，它是整个人口 Y_j 的真正值（该值与样本的 Y 的观测值不同，因为有量度误差）与 Y_j 的估计值 $\hat{Y}_j$ 的差别（Berry，1993）。在线性回归中，如果要将样本的结果推论到整体人口，误差有某些必要的假设，包括平均数为零、常数方差或同方差、正态分布、误差间无相关和误差与自变量不相关。

有时我们会利用残差去验证这些假设，e_j 作为ϵ_j的估计值。当违反某些假设时（零平均值、正态分布），问题相对较小，但违反其他假设便是大问题。异方差会增加标准差并导致检验的统计意义不准确，以及本身可能出现非相加或非线性的特征。与自变量相关，误差变项通常会显示错误的指定，这会使统计推论出现偏差、低效率或不准确的现象。

在线性回归中，残差直接由回归方程计算出来。logistic 回归是根据不同的分析概念产生不同的层次（概率、发生比、logit）的残差。logistic 回归的残差分

析的主要目的是要确定该个案在哪个模型中表现不佳，或对哪个模型的估计参数影响较多。

观测和预测概率之差是 $e_j = P(Y_j = 1) - \hat{P}(Y_j = 1)$，其中 $P(Y_j = 1)$ 是模型 $Y = 1$ 的估计。正如霍斯默和莱默苏所解释的，在线性回归中，我们可以假设误差与 Y 条件平均值相互独立，但在 logistic 回归中，误差方差是条件平均值的一个函数。为此，残差(估计误差)的标准化要通过调整他们的标准误差。皮尔逊(Hosmer & Lemeshow，1989)或标准化(SPSS)或卡方(SAS)的残差是：

$$r_j = z_j = x_j = \frac{P(Y_j = 1) - \hat{P}(Y_j = 1)}{\sqrt{\hat{P}(Y_j = 1)[1 - \hat{P}(Y_j = 1)]}}$$

这只是观测和估计概率的差别除以估计概率的二项标准偏差。对于大样本，标准残余 z_j 应该是平均值为 0 和标准偏差为 1 的正态分布。z_j 大的正或负值表示，个案 j 的模型拟合度很差。因为 z_j 应该是正态分布，95%的案件的数值应在−2 和 2 之间，99%的案件的数值应在−2.5 和+2.5 之间。

一种替代或补充皮尔逊残差的是偏离残差，它等于 $d_j = -2\ \ln$ (正确组别的预测概率)。偏离残差指的是每个个案对 D_M 的贡献。如 z_j，对于大样本，d_j 应是平均数为 0 和标准偏差为 1 的正态分布。第三个残差，残差 logit，等于残差 e_j 除以其方差(而非其标准残差的标准偏差)，写成：

$$l_j = \frac{P(Y_j = 1) - \hat{P}(Y_j = 1)}{\hat{P}(Y_j = 1)[1 - \hat{P}(Y_j = 1)]}$$

残差的非正态性

OLS 回归通常假设误差为正态分布。在小样本中，如果违反这一假设，则基于回归方程(例如回归系数的显著性)，它呈现出不准确的统计推断。但在大样本中，不准确的统计推断被认为是不重要的，因为根据中心极限定理，它表示重复抽样足够大的样本所得出的回归系数会呈正态分布，其平均值(相当于整体人口的平均数)和方差是已知的。在 logistic 回归中，误差没有假设呈正态分布。相反，它假设误差呈二项分布，只在大样本时才接近正态分布。如果残差

是用来估计误差，同时是正态分布（大样本）的，我们就会更有信心地认为，我们的统计推断是正确的，因为正态（我们正考虑的分布）和二项（假定分布）在大样本下接近相同。不同于线性回归分析，如果我们发现小样本的残差不是正态分布的，就不必太担心有关统计推断的效度。我们可以用 SPSS（1999b）将标准或偏离残差对正态曲线做图，或正态概率图。更重要的是，我们可以使用标准残差和偏离残差去确定个案的模型拟合不佳，正或负的标准或偏离残差的绝对值大于 2 的个案[19]，可以帮助我们确定对模型拟合有不良影响，或对模型参数估计有巨大影响的个案。

侦察和处理有影响力的个案

我们可以通过杠杆统计（leverage statistics）的高数值，或最大值 h_i，把对 logistic 回归模型中的参数有潜在巨大影响的个案找出来[20]。在线性回归中，杠杆统计来自方程 $\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jn}Y_k = \sum h_{ij}Y_i$，它代表个案 j 的 Y 预测值，正如个案 j 和其他个案的 Y 观测值的函数（Fox，1991）。每个系数 h_{ij} 捕捉 Y_i 观测值对预测值 $\hat{Y}_j$ 的影响，可写成 $h_{ii} = \sum (h_{ij})^2$，所以如果我们指定 $h_i = h_{ii}$，就会得到在样本所有个案中，Y_i 对 Y 预测值的整体影响。同样的杠杆统计也衍生于 logistic 回归（Hosmer & Lemeshow，1989：150—151），数值介于 0（没有影响）和 1（完全确定模型的参数）。一个独立方程含有 K 个自变量（包括每个设计变量），或者同等的，在一个公式中，有 k 自由度与 G_M 相关，h_i 的总和等于 $k+1$ 和 h_i 的平均值 $\sum h_i/N = (k+1)/N$。如果个案的最大值大于 $(k+1)/N$，就是高杠杆个案。

其他侦察具有影响力个案的方法包括皮尔逊卡方平方变化和将个案从分析中删除后的 D_M 变化。删除个案 j，皮尔逊 χ^2 的变化是：$\Delta\chi_j^2 = z_j^2/(1-h_j)$，其中 z_j 是个案 j 的标准残差，h_j 是其统计杠杆。D_M 的变化是：$\Delta D_j = d_j^2 - z_j^2 h_j/(1-h_j) = d_j^2 - h_j(\Delta\chi^2)$，其中 d_j 是个案 j 的偏离残差，z_j 是个案 j 的标准残差，h_j 是个案 j 的杠杆统计。

ΔD_j 和 $\Delta_j\chi_j^2$ 是 χ^2 的分布，可借此解释它们的数值。它们各自的平方根应接近正态分布。如果$\sqrt{\Delta D_j}$（SPSS LOGISTIC 回归的学生残差或 SAS PROC LOGISTIC 的 C）或$\sqrt{\Delta\chi_j^2}$小于−2 或大于 2，这显示该个案可能不能很好地拟合模型而需要对个案多加留意。$z_j^2h_j/(1-h_j)$ 的数值表示因为删除个别观测导致回归估计的整体变化，这些数值显示在 SAS PROC LOGISTIC 的可选统计 CBAR 和 SPSS LOGISTIC 回归的 Cook 距离。量度的标准化可以用 $(1-h_j)$ 除以 Cook 距离；$z_j^2h_j/(1-h_j)^2=\text{dbeta}$，这是删除个案 j 导致的回归系数的标准改变。前面所描述的杠杆统计和相关统计是总结个案对模型估计的指针。如果想要更详细的资料，可观察当删除某一个案时，每个系数的变化。SPSS 和 SAS 都把 logistic 回归系数的变化写成 DFBETA。

离群值和残差图

表 4.3 列出残差分析的结果。如果个案的$\sqrt{\Delta D_j}$小于−2 或大于 2 就会检查该个案。该表列出了个案序号、观测和预测值、皮尔逊（ZResid）、学生氏（SResid）和偏离残差（Dev）、杠杆（Lever）和已删除的残差 ΔD_j（DIFDEV）、$\Delta\chi^2$（DIFCHI）和 dbeta（DBETA）。表 4.3 的 A 部分显示表 3.1 模型的残差，B 部分显示删除最极端的离群者，这是图 4.1 分析中两个被确定非线性的其中一个，C 部分显示，从图 4.1 删除两个离群者后的分析结果。

表 4.3 的指标基本上是多余的，皮尔逊 χ^2 的变化 DIFCHI（$\Delta\chi^2$）大约等于皮尔逊残差—ZResid 的平方。偏离残差 Dev（d_j）大约等于学生残差（SResid），同时偏离残差 DIFDEV（ΔD_j）也等于学生残差的平方。基于皮尔逊 χ^2 的残差大于基于 D_M 的残差，但他们所提供的有关个案的基本讯息是相同的。杠杆和 DBETA 提供数据不能做其他诊断。进一步的分析将集中在皮尔逊残差、学生残差、杠杆和 DBETA。

在 A 部分，个案编号 178 最突出，皮尔逊残差是−10.7，学生残差的绝对值大于 3，DBETA 大于 1，所有指针都显示极端不适合。删除该个案，G_M 明显改善了 9.899（$df=1$，$p=0.003$）和 R_L^2 增加 0.034。显然，删除该个案，模型拟

表 4.3 logistic 回归诊断概要

个案	PMRJ5 的观测值	Pred	ZResid	Dev	SResid	Lever	DIFCHI	DIFDEV	DBETA
				A. 全模型					
66	1	0.0991	3.0143	2.1500	2.1637	0.0127	9.20	4.68	0.12
94	0	0.8608	−2.4864	−1.9858	−2.0325	0.0455	6.48	4.13	0.31
139	1	0.0815	3.3565	2.2391	2.2668	0.0243	11.55	5.14	0.29
148	1	0.0612	3.9183	2.3641	2.3762	0.0102	15.51	5.65	0.16
178	0	0.9914	−10.7055	−3.0823	−3.0983	0.0103	115.80	9.60	1.21
201	1	0.0650	3.7937	2.3383	2.3526	0.0121	14.57	5.53	0.18
$G_M = 108.257$	5df	$p = 0.0000$	$R_L^2 = 0.367$						
				B. 删除最极端的个案					
1	0	0.9122	−3.2239	−2.2059	−2.2557	0.0436	10.87	5.09	0.50
66	1	0.0894	3.1913	2.1975	2.2116	0.0127	10.32	4.89	0.13
94	0	0.8786	−2.6903	−2.0536	−2.0999	0.0436	7.57	4.41	0.34
139	1	0.08613	0.2577	2.2146	2.2444	0.0264	10.90	5.04	0.30
148	1	0.05104	0.3137	2.4396	2.4515	0.0097	18.79	6.01	0.18
200	1	0.06613	0.7601	2.3312	2.3463	0.0129	14.32	5.51	0.19
$G_M = 118.156$	5df	$p = 0.0000$	$R_L^2 = 0.401$						
				C. 删除图 4.1 的离群个案					
66	1	0.0808	3.3732	2.2432	2.2573	0.0125	11.52	5.10	0.15
94	0	0.8858	−2.7852	−2.0832	−2.1289	0.0425	8.10	4.53	0.36
133	0	0.8675	−2.5584	−2.0105	−2.0506	0.0387	6.81	4.20	0.27
139	1	0.0885	3.2099	2.2023	2.2332	0.0275	10.59	4.99	0.30
148	1	0.0441	4.6536	2.4982	2.5097	0.0092	21.86	6.30	0.20
200	1	0.0675	3.7175	2.3221	2.3378	0.0135	14.01	5.47	0.19
$G_M = 122.634$	5df	$p = 0.0000$	$R_L^2 = 0.416$						

注:列出所有学生残差大于 2.0000000 的个案。

合更好。在 B 部分,删除个案 178 后,没有其他个案这样明显地偏离。个案 148 有最高的皮尔逊残差和最高学生残差,但把它删除后,对 logistic 回归系数的影响不大。个案 1 对 logistic 回归系数 (DBETA = 0.50) 的影响更大,而且它的

皮尔逊残差程度位居第四,学生氏残差位居第三,六个个案的偏离残差被选定为离群。个案 1 还有另一个附加特点,就是图 4.1 确定的两个离群者的一个有非线性。把个案 1 删掉,G_M 提高了 4.478($df = 1$, $p = 0.038$) ,R_L^2 增加了 0.015,不过删掉个案 1 的结果改善幅度比起删掉个案 178 少。

个案 1 和个案 178 应该从分析中删除吗?答案是要再做更深入的检验。这两个个案都是白人,一男一女,所报告的违法观念较低,也没有使用大麻或毒品,而且很少喝酒。个案 1(女)比个案 178(男)的违法观念稍高,接触违法朋友的水平也较低,因此比个案 178 更不符合该模型。虽然很特别,但结果都可信,所以两种个案都应保留下来。这也许对扩展该模型有帮助,模型该包括哪些变量,才能解释为什么一个人认为违法没有大错,但自己又不喝酒、吸食大麻或其他非法毒品呢?

休梅克(Shoemaker)等人(1984)以及霍斯默和莱默苏(1989)都认为可应用图形技术诊断 logistic 回归,视觉比起数字更直观、更有吸引力。例如霍斯默和莱默苏(1989)建议将 DIFCHI, DIFDEV 和 DBETA 与预测值做图,以侦测离群个案。图 4.3 的第一栏的图表就是一个例子,每个图中都有两条曲线,一个从左至右下滑(个案的 PMRJ5 观测值是 1),另一个从左至右增加(个案的 PMRJ5 观测值是 0)。图左上角和右上角的个案的模型拟合较差。从 χ^2 变化(DIFCHI)对预测值的图像看到,有一个个案是极端的离群者,它的 DIFCHI 大于 100。从表 4.3 可以看到,这就是个案 178。同样的,个案 178 的 DIFDEV 图,显示它的 DIFDEV 值大于 8, DBETA 值大于 1。图 4.3 第一栏的图看到其他个案似乎都相当集中。不过,当删除个案 178 后,图表改变了,又出现其他离群个案。

图 4.3 第二栏也是删除个案 178 后的图像,比第一栏更分散,因为量表不同(DIFCHI 从 0—100 变为 0—20, DIFDEV 从 0—10 变为 0—6.25, DBETA 从 0—1.25 变成 0.5)。把 PMRJ5 = 1 和 PMRJ5 = 0 的两条曲线结合起来,形成一个酒杯状形的图案(特别是 DIFDEV),所有的离群者都在右上角和左上角。相对于第一栏,图的上部分与其余部分有相对平稳的过渡。最离群的个案也不是像第一栏的个案 178 那样明显地分开。这反映在表4.3B 中,没有一个学生氏残差的绝对值大于 2.5,或 DBETA 大于 1。

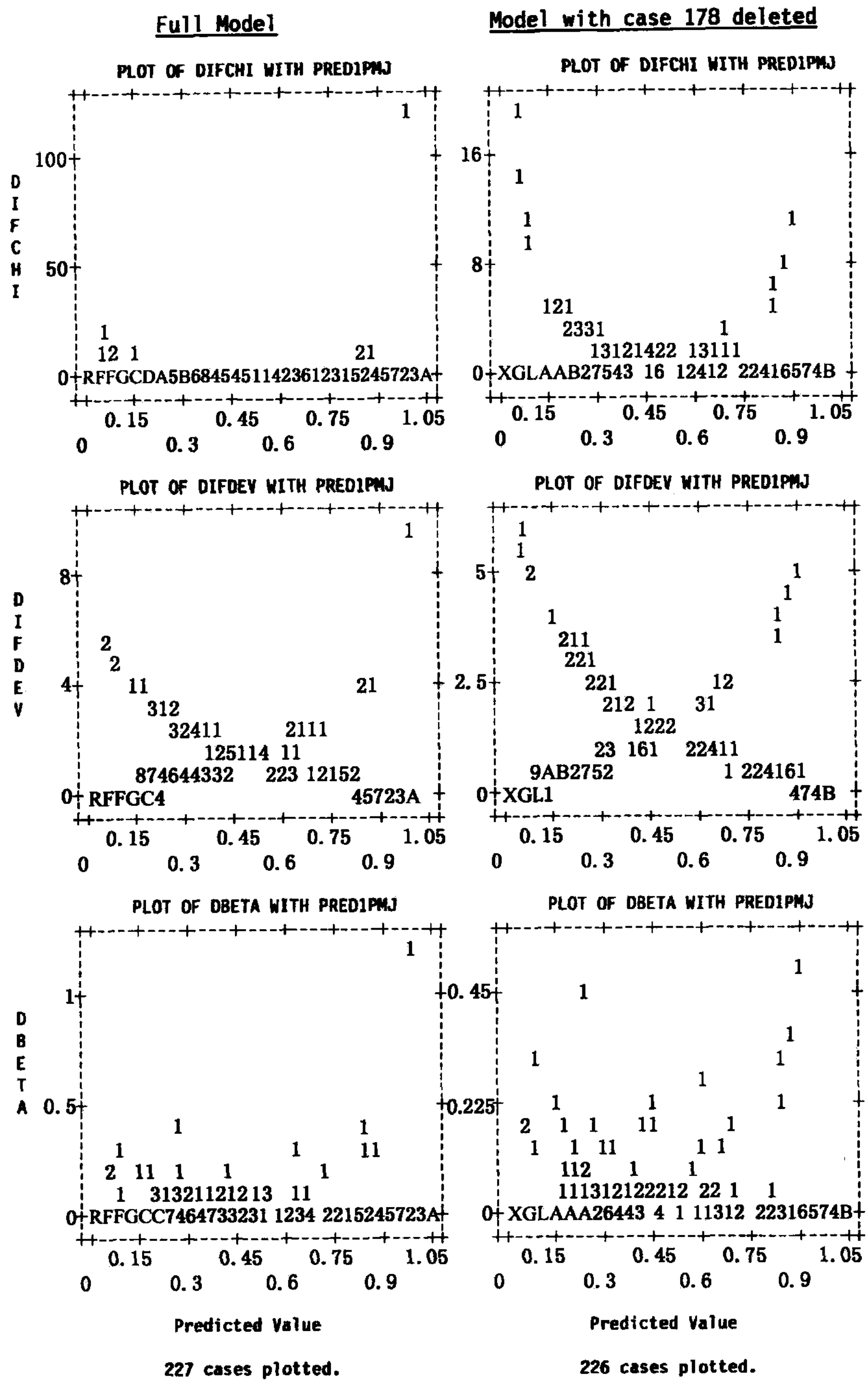


图 4.3 logistic 回归诊断图

过度分散和过度集中[21]

假设 logistic 回归的误差是二元误差，方差 $\sigma_Y^2 = P_{Y=1}(1-P_{Y=1})$，当数据是个体层次，即每个共变形式只有一个个案，那就符合这个假设。但对于组群数据，个案集合在一起，共变形式便被视为“个案”，这可能违反该假设。原因有可能遗漏了一个重要的预测变量——样本分层或潜在人口分布与假定分布不同。过度分散指个案的 σ_Y^2 大于预期，过度集中指个案的 σ_Y^2 少于预期。对于组群数据，我们计算 $\delta = D/df$，其中 D 是共变形式的偏离统计(deviant statistics)(例如 SPSS NOMREG 的拟合表会有显示)，df 是与偏离统计关联的自由度(也在同一个表上)，$\delta > 1$ 显示过度分散，$\delta < 1$ 表示过度集中。过度分散和过度集中都会得出不正确的标准误差而导致不正确的统计推断。调整标准误差的方法是乘以 δ 的平方根：调整标准差 = 标准误 $\times\sqrt{\delta}$。SPSS NOMREG 的方法是利用皮尔逊或偏离 χ^2，或用户指定的 δ 值(SPSS NOMREG 的 N)。注意，图 3.2 的偏离 $\chi^2 = 158.266$，$df = 151$，$\delta = 158.266/151 = 1.05$，表示很少或根本没有过度分散或过度集中。

logistic 回归诊断的规程

现有的回归软件能快速而简易地检验 logistic 回归分析中的共线性，因此能在分析前侦察出潜在的问题。博克斯—蒂德韦尔检验是一个快速而简单的非线性检验方法，而且它对稍微的偏离性不会过于敏感，应该纳入 logistic 回归的标准程序。是否去检验非相加性，这要看有没有理论或其他理由认为模型存在交互项。模型的非线性和非相加性的处理都要略为谨慎，因为真正的危险是过度的模型拟合，可能只捕捉了随机变化，而不是系统化的行为规律。

利用 logistic 回归诊断，就像线性回归诊断，是一门艺术而非科学。诊断统计显示出潜在的问题，需要进一步研究那些问题是什么，是否需要采取补救行

动，需要仔细观测不寻常个案后才决定。在 200 至 250 个个案样本中，单纯随机抽样的变异就能产生 10 到 12 个个案的正态分布变量如偏离残差或学生残差的标准值大于 2 或小于−2。即使个案有非常大的残差，如表 4.3 的个案 178，也不一定指模型有问题，这种模型不可能太明确，因为人的自由意志和个人选择肯定会影响人类行为的完美预测。

一般而言，我们应当对模型进行一些诊断，以检验错误编号的数据和模型概念的弱点。诊断分析至少包括学生残差、杠杆和 dbeta。学生残差小于−3 和大于 3 都要仔细检查；小于−2 或大于 2 可能要留意。皮尔逊残差的缺点是，它提供的信息往往与学生残差和偏离残差（不是皮尔逊残差）重复，这些偏离残差是模型参数估计的准则，因此，它们应是更合适的残差分析。但皮尔逊残差的优点在于它的数值比偏离或学生残差大，尤其是离群者，因此，能容易地找出离群者（像表 4.3A 部分的个案 178）。杠杆值与 $(k+1)/N$ 是预期值的几倍（在这个例子中约 $5/227=0.02$），要密切注意。dbeta 的数值大，特别是当数值大于 1 时（记住，这是一个标准化的测量），也要仔细研究。是否以图像视觉显示，这是个人喜好。最重要的是，诊断出来的极端值需要再仔细检测，并重新考虑模型。

第 5 章 | 多分类 logistic 回归及其替代方法

logistic 回归分析可由二分变量延伸到多分类(有两个或两个以上的类别)的名义或定序因变量。在 logistic 回归的文献中,这种模型称为多分类或多项 logistic 回归模型。这里,二分和多分类指 logistic 回归模型,二项和多项指的是对数模型。由于对数模型可衍生出多分类的 logistic 回归模型,因此多分类因变量的 logistic 回归模型可算是多项对数模型的一个特殊例子(Agresti, 1990; Aldrich & Nelson, 1984, DeMaris, 1992; Knoke & Burke, 1980)。

从数学角度来说,二分 logistic 回归模型延伸至多分类因变量是很简单的。因变量的其中一个值(通常是第一个或最后一个)被指定为参考类,$Y = h_0$,然后比较落在另一组或参考类别的概率。对于名义变量,这可能是直接的比较,类似 logistic 回归模型中的二分自变量的对比指标。但对定序变量,对比连续类别,方法如同二分 logistic 回归模型自变量的重复或赫尔默特对比。

因变量具有 M 个类别,这需要计算 $M-1$ 个方程,每个类别相对于参考类别,以描述因变量和自变量之间的关系。除了参考类,因变量的每个类别,我们可以写成:

$$g_h(X_1, X_2, \cdots, X_k) = e^{(a_h+b_{h1}X_1+b_{h2}X_2+\cdots+b_{hk}X_k)},$$

$$h = 1, 2, \cdots, M-1 \qquad [5.1]$$

其中下标 k 是指特定的 X 自变量,下标 h 指 Y 因变量的特定值,参考类别 $g_0(X_1, X_2, \cdots, X_k) = 1$。$Y$ 等于 h 的任何值,除 h_0 外的概率等于:

$$P(Y = h \mid X_1, X_2, \cdots, X_k) = \frac{e^{(a_h+b_{h1}X_1+b_{h2}X_2+\cdots+b_{hk}X_k)}}{1+\sum_{h=1}^{M-1} e^{(a_h+b_{h1}X_1+b_{h2}X_2+\cdots+b_{hk}X_k)}},$$

$$h = 1, 2, \cdots, M-1 \qquad [5.2]$$

不包括类别 $h_0 = M$ 或 0，

$$P(Y = h_0 \mid X_1, X_2, \cdots, X_k) = \frac{1}{1 + \sum_{h=1}^{M-1} e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}}$$

$$h = 1, 2, \cdots, M-1 \qquad [5.3]$$

请注意，当 $M = 2$，就是二分因变量的 logistic 回归模型，参考类别是第一类，$h_0 = 0$，总共 $M-1 = 1$ 个方程来描述变量间的关系。多分类名义因变量的 logistic 回归模型可用 SAS CATMOD 或 SPSS 10(之前版本用 LOGLINEAR)来计算，这两个软件使用对数线性分析程序去计算多分类 logistic 回归模型，过程相当繁琐。SPSS 10 的 NOMREG 提供更简易的名义因变量 logistic 回归模型方法。SAS LOSISTIC 和 SPSS PLUM 同样具有简易的多项定序因变量分析。虽然这本书的重点是 SAS 和 SPSS，值得一提的还有 STATA(1999)，它具有范围广泛的 logistic 回归程序，包括用在名义因变量的 MLOGIT 和定序因变量的 OLOGIT。

为了说明多分类 logistic 回归，我们把前面例子的因变量大麻使用率，改成具有四个类别的药物使用者：

1. 不使用者：报告他们在过去一年没有喝酒，或吸食大麻、海洛因、可卡因、安非他明、巴比妥或迷幻丸。

2. 饮酒者：报告他们在过去一年喝过酒，但不吸食任何违禁药物。

3. 吸食大麻者：报告他们吸食大麻，同时，除一个个案之外，全部喝过酒。

4. 吸食多种违禁药物的人：报告他们使用一种或多种的“硬性”毒品(海洛因、可卡因、安非他明、巴比妥、迷幻丸)。除一个个案之外(只报称吸食一次或一种毒品)，他们还会喝酒和吸食大麻。

根据法律后果的程度，从最不严重到最严重的毒品，这些变量可被合理地当做定序变量。或者，如果没有法律后果的话，可视为名义变量，不过这样会有点争议性，因此，分类或定序模型都可考虑。但之前的模型需要做一个额外的修改，因为因变量有四个类别，种族的“其他”类别的个案太少，所以要重新编码

分成白人和非白人，否则会出现格数值为零的问题和不稳定的系数估计及标准误差。

多分类名义因变量

图 5.1 是 SPSS NOMREG[22] 的输出，DRGTYPE 是因变量，用 DRGTYPE 的对比去比较(a)不喝酒的人；(b)不吸食大麻的人；(c)不吸食多种药品的人，得出 $g_1(X)$、$g_2(X)$和 $g_3(X)$的三个函数，定义为：

g_1 = logit (喝一点酒对不吸食药物的概率)；

g_2 = logit (吸食大麻对不吸食药物的概率)；

g_3 = logit (使用违禁药物对不吸食药物的概率)。

在图 5.1 中，g_1，g_2，g_3 的方程都使用非标准系数，

$g_1 = 0.165(\text{EDF5}) - 0.271(\text{BELIEF4}) + 0.505(\text{SEX}) + 1.616(\text{WHITE}) + 5.085$；

$g_2 = 0.506(\text{EDF5}) - 0.285(\text{BELIEF4}) - 0.920(\text{SEX}) + 0.357(\text{WHITE}) + 2.503$；

$g_3 = 0.663(\text{EDF5}) - 0.360(\text{BELIEF4}) - 2.224(\text{SEX}) + 2.209(\text{WHITE}) + 0.768$。

g_1、g_2 和 g_3 每个 logistic 函数的 R^2、η^2 和标准 logistic 回归系数都要分开计算(类似判别分析的线性判别函数的典型相关和标准判别函数系数，见 Klecka，1980)。全模型 R^2 的计算是基于预测概率和四个类别的观测分类。预测表已包括在 SPSS NOMREG，或在 SAS CATMOD 中，它们应用方程式 5.2 和 5.3 计算 Y 每个值的分类概率，包括参考类别，然后把个案分到最高概率的类别内，SAS PROC FREQ 会有这个分类表格。

一旦做好了分类表格，预测指数如图 5.1 的分类表那样，根据第二章的方法计算出来。多项名义因变量模型的 λ_p 与 τ_p 的差别与其他提及过的预测效率不同，更有理有据。

图 5.1 的模型相当不错，模型 χ^2 有显著性，McFadden R_L^2 等于 0.28。

SPSS NOMREG DRGTYPE5 与性别种族和 EDF5 BLIEF4/模型＝EDF5 BELIEF4 性别种族/打印＝拟合 IRT 参数概要分类表/量度＝偏离
注意:437(70.9%)的格数值是 0。

		N
DRGTYP5	1.000 alcohol	87
	2.000 marijuana	50
	3.000 drugs	31
	4.000 nonuser	59
SEX	1 male	110
	2 female	117
ETHN	1 white	175
	2 nonwhite	52
Valid		227
Missing		30
Total		257

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	549.126			
Final	379.778	169.348	12	0.000

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	379.778	0.000	0	.
EDF5	444.950	65.172	3	0.000
BELIEF4	396.795	17.017	3	0.001
SEX	404.765	24.987	3	0.000
ETHN	399.995	20.217	3	0.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	479.072	447	0.142
Deviance	341.094	447	1.000

Pseudo R-Square

Cox and Snell	0.526
Nagelkerke	0.566
McFadden	0.282

R^2 = 0.303
R_1^2 = 0.189
R_2^2 = 0.149
R_3^3 = 0.337
λ_p = 0.300 p =0.000
τ_p = 0.399 p =0.000

Classification

	Predicted				
Observed	1.000 alcohol	2.000 marijuana	3.000 drugs	4.000 nonuser	Percent Correct
1.000 alcohol	61	7	2	17	70.1%
2.000 marijuana	21	16	8	5	32.0%
3.000 drugs	6	7	18	0	58.1%
4.000 nonuser	20	5	0	34	57.6%
Overall Percentage	47.6%	15.4%	12.3%	24.7%	56.8%

图 5.1 多分类名义 logistic 回归

logit(*Y*)的解释方差应因变量的类别而变化，最高是 g_3(使用多种药物的人)，最低是 g_2(吸食大麻)。总体模型，如似然比检验表显示，所有四个预测变量都是显著的。正如图 5.1 顶部的 SPSS NOMREG 所示，偏离 χ^2 已修正了分散的问

Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)		Standardized Logistic Regression Coefficients
DRGTYP5								Lower Bound	Upper Bound	Note: These have been edited into the output.
1.000 alcohol	Intercept	5.085	2.463	4.264	1	0.039				
	EDF5	0.165	0.091	3.303	1	0.069	1.179	0.987	1.409	0.209
	BELIEF4	-0.271	0.070	14.906	1	0.000	0.763	0.665	0.875	-0.319
	[SEX=1]	0.505	0.338	2.226	1	0.136	1.656	0.854	3.214	0.075
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	1.616	0.400	16.277	1	0.000	5.032	2.295	11.031	0.202
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
2.000 marijuana	Intercept	2.503	2.674	0.876	1	0.349				
	EDF5	0.506	0.096	27.544	1	0.000	1.659	1.373	2.004	0.671
	BELIEF4	-0.285	0.078	13.319	1	0.000	0.752	0.645	0.876	-0.350
	[SEX=1]	-0.920	0.439	4.401	1	0.036	0.398	0.169	0.941	-0.143
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	0.357	0.462	0.596	1	0.440	1.428	0.578	3.531	0.047
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
3.000 drugs	Intercept	0.768	3.049	0.064	1	0.801				
	EDF5	0.633	0.106	35.976	1	0.000	1.883	1.531	2.316	0.677
	BELIEF4	-0.360	0.086	17.515	1	0.000	0.698	0.590	0.826	-0.357
	[SEX=1]	-2.224	0.619	12.893	1	0.000	0.108	3.211E-02	0.364	-0.279
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	2.209	0.841	6.901	1	0.009	9.104	1.752	47.3 2	0.233
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	

a This parameter is set to zero because it is redundant.

图 5.1 多分类名义 logistic 回归(续)

题,这是因为偏离 χ^2 似乎有点低于自由度($\chi^2 = 341$, df $= 447$, $\chi^2/\mathrm{df} = 0.76$),表示过度集中。调整分散会影响沃尔德系数的显著性。喝酒的人标准化系数(不在 SPSS 的输出里,但可像 λ_p、τ_p 和 R^2 那样加到 SPSS 的输出中)表明,最好的预测者是违法观念——认为违法是错误的,其次是种族(白人受访者

比非白人更容易喝酒)。接触违法朋友的沃尔德统计是临界($p=0.069$),性别没有显著性。对吸食大麻和多种药物的预测,最好的预测者是接触违法朋友,其次是观念和性别。种族对吸食大麻不具显著性,但白人受访者比非白人更有可能食用多种药物。根据图 5.1 的分类表,预测效率的指数 $\lambda_{\mathrm{p}}=0.300$ 和 $\tau_{\mathrm{p}}=0.399$ 均有显著性,而且强度也不弱。

多分类或多项式定序因变量

当因变量是定序量表,就会有很多可能的分析方法,绝不仅限于 logistic 回归分析。更详细的讨论,见 Agresti(1990:318—332);Long(1997:114—147);Clogg 和 Shihadeh(1994)。简单来说,可行方法包括:

1. 忽略因变量分类的顺序而把它当做名义变量;
2. 权当变量是真正的定序变量;
3. 认为变量的衡量是定序量度,但只代表定距或比率的粗略量度;
4. 将变量视为定距变量。

第一个方法的例子是多项对数或 logistic 回归模型,因变量是名义变量,如图 5.1,也可以用判别分析(Klecka, 1980)。第二个方法的例子是使用累积对数模型,因变量每类的转换不仅与参考类别相比,还与所有更高(或更低)类别做比较。第三个方法假设是潜在的定距量表,那就可用 LISREL 的加权最小二乘法(WLS)分析得出多项相关(Joreskog & Sörbom, 1988)[23]。第四个方法可采用定序因变量的最小二乘回归。

选择哪个方法需要小心判断。第四个方法能有效地假定数据的测量比实际上更精确,对有大量分类的定序变量可能是合理的。应用 WLS 估计多项相关似乎是一个比较好的选择,因为它适用于不同分类的定序变量。假设测定量表是连续的(政治保守主义、药物的严重程度)本质上是可行的。第四和第三个方法都允许预测值落在观测值的范围之外,在不精确测量的假设下,这可能是合理的。

本书不建议使用现成软件程序,例如 SAS PROC LOGISTIC 和 SPSS

PLUM(可计算定序因变量的分类 logistic 回归模型),但它们都是使用因变量的累积对数模型。这种模型假设三个方程中每个自变量的系数都不变,即 $b_{EDF5,1} = b_{EDF5,2} = b_{EDF5,3}$; $b_{SEX,1} = b_{SEX,2} = b_{SEX,3}$ 等等(平行斜率),其中,下标指的是该系数的自变量,数字是系数出现的方程式(1, 2 或 3)。平行斜率模型的方程,只有拦截不同,不同组别的自变量的效应都被假定是固定的。虽然 SAS PROC LOGISTIC 或 SPSS PLUM 都很容易算出模型,但这未必是最合适的模型。

图 5.2 显示了药物用户类型 DRGTYPE 的分析结果,这些变量在 SAS PROC LOGISTIC 是定序变量。SAS 假设斜率相等。图 5.2 原假设的分数检验显示斜率等于 32.066,自由度为 8,在 0.0001 水平有显著性。由于分数检验有显著性,平行斜率的假设被拒绝,这表示模型不假定平行斜率会较为恰当。图 5.1 显示了拒绝相等斜率模型的原因就是各自变量的效应强度和显著性都不同:EDF5(只喝酒的人与不使用者比较没有显著性),性别(不使用者与只喝酒比较没有显著性;但严重滥用药物的人比大麻使用者强),ETHN(吸食大麻与不使用者比较没有显著性)。图 5.1 的模型的各系数差异(特别种族系数向下向上的格局)表示,把 DRGTYPE 当做名义变量可能是最好的选择。

SPSS PLUM 与 SAS LOGISTIC 提供的信息大致相同,但 SPSS PLUM 不包括图 5.2 底部的信息(预测概率与观测回应的关系),也不包括 SPSS NOMREG 中的皮尔逊和偏离拟合 χ^2 统计以及 McFadden R_L^2,后者(包括整体模型和每个各自函数的 R^2)已被编辑过才出现在图 5.2 的 SAS 输出中。无论是正态分布、正或负偏斜还是有许多极端值的因变量,SPSS PLUM 都有不同的方法去处理它的对数分布。运用 SPSS PLUM 和 SAS LOGISTIC,首先要储存预测值,然后利用预测值和观测值列出列联表(SAS PROC FREQ 或 SPSS CROSSTABS)去分析分类的准确度。图 5.2 的例子做同样的分析,结果得出 $\lambda_p = 0.229(p = 0.000)$, $\tau_p = 0.208(p = 0.000)$,两者都比图 5.1 少,再次说明最好把因变量当做名义变量,而不是定序变量。但总的来说,定序变量如图 5.2 底部的统计如 Gamma 和 Tau-a 所提供的信息会比 λ_p 和 τ_p 多,因为前两者不像后者会把因变量类别的排列信息纳入分析中。

```
data;
  infile 'saslr16a.dat' missover linesize=60 firstobs=1 obs=257;
  input ID F66 SEX 8 ETHN 10 USR5 12 PDRUGS5 14-15 PMRJ5 17-18 PALC5 20-21
        DRGTYP5 23-24 EDP5 26-33 BELIEF4 35-42 MEANSCIN 44-51 MEANFAIN 53-60;
  if ethn=1 then white=1; if ethn=2 then white=0; if ethn=3 then white=0;
  if drgtyp5=1 then drgtyp5r=4; if drgtyp5=2 then drgtyp5r=3;
  if drgtyp5=3 then drgtyp5r=2; if drgtyp5=4 then drgtyp5r=1;const5=1;
run;
proc logistic;
  model drgtyp5r=edp5 belief4 sex white;run;
```

Data Set: WORK.DATA1
Response Variable: DRGTYP5R
Response Levels: 4
Number of Observations: 227
Link Function: Logit

Response Profile

Ordered Value	DRGTYP5R	Count
1	1	31
2	2	50
3	3	87
4	4	59

WARNING: 30 observation(s) were deleted due to missing values for the response or explanatory variables.

Score Test for the Proportional Odds Assumption
Chi-Square = 32.0660 with 8 DF (p=0.0001)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates	
		.		R^2_1 = 0.004
				R^2_2 = 0.089
AIC	606.600	485.626	.	R^2_3 = 0.151
SC	616.875	509.600	.	
-2 LOG L	600.600	471.626	128.975 with 4 DF (p=0.0001)	R^2_0 = 0.264
Score	.	.	97.395 with 4 DF (p=0.0001)	R_L^2 = 0.215

RSquare = 0.433　　　　Adjusted RSquare = 0.467

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate b_{sas}	Odds Ratio	Standardized coefficient $b^*=(b)(s_X)/s_Y$
INTERCP1	1	-1.3616	1.4611	0.8684	0.3514	.	.	.
INTERCP2	1	0.6157	1.4513	0.1800	0.6714	.	.	.
INTERCP3	1	2.9884	1.4655	4.1583	0.0414	.	.	.
EDP5	1	0.2701	0.0424	40.5402	0.0001	0.633781	1.310	0.343
BELIEF4	1	-0.1774	0.0426	17.3225	0.0001	-0.386429	0.837	- 0.2 9
SEX	1	-0.7905	0.2630	9.0312	0.0027	-0.218288	0.454	- 0.118
WHITE	1	0.8343	0.3167	6.9391	0.0084	0.193729	2.303	0.105

Association of Predicted Probabilities and Observed Responses

Concordant	= 80.5%	Somers' D	= 0.623
Discordant	= 18.2%	Gamma	= 0.631
Tied	= 1.3%	Tau-a	= 0.449
(18509 pairs)		c	= 0.811

图 5.2　定序 logistic 回归的 SAS 输出结果

结论

多项因变量 logistic 回归分析的重点是模型整体上是否合适。未涉及的定

序因变量 logistic 回归模型的问题包括：因变量的预测值超出应有的范围和异方差，这都会促使模型的发展。其他模型可能比 logistic 回归更适合，但取决于因变量的潜在量表的假设和因变量与自变量的函数形式（线性、单向、非单向）。如果是定距的量度，关系是线性或单向，多元加权最小二乘可能是最好的选择。非单调的关系，尤其是类别比较少的因变量，最适当的方法可能是把因变量当做名义变量。当因变量是名义，或者可视为名义变量的定序变量（少类别），判别分析是一个值得考虑的方法（Klecka，1980）。另外，排除 logistic 回归（Bess & Grey，1984；Hosmer & Lemeshow，1989：230—232），因为它得出的结果似乎与多项对数/多分类 logistic 回归模型不充分一致，所以不建议采用。除非没有多分类 logistic 回归分析软件（这种情况越来越少，主要统计软件都已包括多分类 logistic 回归分析），这种方法才有优势。

由于易用、灵活和普及程度高，目前 logistic 回归分析很容易被滥用。滥用 logistic 回归不会比滥用线性回归或任何一样技术更有好处。重要的是明白这个方法的弱点和优点，logistic 回归特别适合分析二分和无序名义多分类的因变量。多分类定序因变量可能使用多分类 logistic 回归分析。但其他模型包括线性回归和多项相关加权最小二乘就要认真考虑。多分类定序变量的因变量从技术性层面来说，使用 logistic 回归是最差的方法，其他分析方法可能比 logistic 回归好。但是鉴于使用方便、灵活和普及程度高（特别是 logistic 回归软件不断地改良），尤其当 OLS 不适用时，logistic 回归可用于分析不同类型的因变量。

附 录 | 概率

估计事件发生的概率是利用它在人口或样本中的相对频率。例如，如果 $n_{Y=1}$ 是样本 $Y=1$ 的个案数量，N 是整个样本的总数，那么：

1. Y 等于 1 的概率写为 $\mathrm{P}(Y=1)$；

2. $\mathrm{P}(Y=1)=n_{Y=1}/N$；

3. Y 不等于 1 的概率是 $\mathrm{P}(Y\neq 1)=1-\mathrm{P}(Y=1)=1-(n_{Y=1}/N)=(N-n_{Y=1})/N$；

4. 概率的最低可能为 0 ($n_{Y=1}=0$ 暗中 $n_{Y=1}/N=0$)；

5. 概率的最高可能为 1 ($n_{Y=1}=N$ 暗中 $n_{Y=1}/N=1$)。

两个独立事件的联合概率(事件发生是互不相关)就是这两个概率的乘数。例如，X 和 Y 都等于 1 的概率，如果 X 和 Y 是不相关的，$\mathrm{P}(Y=1$ 和 $X=1)=\mathrm{P}(Y=1)\times\mathrm{P}(X=1)$。如果 X 和 Y 是相关的(例如，Y 等于 1 的概率是取决于 X 值)，则 $\mathrm{P}(Y=1$ 和 $X=1)$ 将不等于 $\mathrm{P}(Y=1)\times\mathrm{P}(X=1)$。相反，我们要考虑当 $X=1$ 时 $Y=1$ 的条件概率，或 $\mathrm{P}(Y=1\mid X=1)$。

$Y=1$ 的条件概率是当某些变量有特定值时，$Y=1$ 的概率，[有时指 $\mathrm{P}(Y=1)$，无论其他变量为何值，就如 $Y=1$ 的无条件概率]。例如图 2.1，使用大麻普及率等于 1 的概率是 $\mathrm{P}(\mathrm{PMRJ5}=1)=0.35$ (男性和女性总和；详细的数据不显示)。使用大麻普及率等于 1 的条件概率是女性为 $\mathrm{P}(\mathrm{PMRJ5}=1\mid \mathrm{SEX}=0)=0.45$，男性为 $\mathrm{P}(\mathrm{PMRJ5}=1\mid \mathrm{SEX}=1)=0.25$。对于一个二分变量，编码为 0 或 1，该变量的概率等于 1 等于变量的平均数，变量等于 1 的条件概率等于该变量的条件平均值。

注释

[1] 虽然被建模的关系往往代表着因果关系，单一变量的预测受到一个或多个预测变量的影响，但并非总是如此。我们可以很容易地从结果预测出原因（例如，从收入预测那人到底是男性还是女性），或从原因预测出结果（从男或女预测到他/她的收入）。这本书的重点在于预测，而不是因果关系，虽然表述上有时用因果关系、自变量、因变量、后果或预测变量，但不一定意味着因果关系。所有关系绝对是预测性的，可能本身有因果关系。

[2] 数据来自全国青年调查，样本是 1976 年全国家庭概率抽样的 11 至 17 岁青少年和 1992 年的 27 至 33 岁青年。数据是从 1976 年至 1980 年每年收集一次，然后 1983 年至 1992 年每隔 3 年收集一次。数据包括自报的违法行为、家庭关系、学校表现以及社会人口特征。详细的抽样和量表请参看埃利奥特（1985，1989）。现在只用 1980 年 16 岁青少年的数据。散点图的数字和符号指个案数目：1 表示一个个案，2 表示两个，9 表示 9 个；字母 A 到 Z，A 代表 10 个个案，B 代表 11 个，…，Z 代表 35 个。当超过 35 时，用星号（*）代表。

[3] 详尽的量表水平，请参考 Agresti 和 Finlay（1997：12—17）。

[4] Y 无条件平均值就是 $\overline{Y} = \sum$。Y 的条件平均值（特定 X 值）是只计算有 X 值的个案，$\overline{Y}_{X=i} = \sum Y_j$，其中 i 是 X 值，Y_{ij} 是当 $X = i$，个案 $j = 1, 2, \cdots, n_i$ 的数值。

[5] 附录部分有简短的概率讨论，包括条件概率。

[6] 贝里和费尔德曼（1985：63—64）以及刘易斯—贝克（1980：44）讨论到对数变换解决变量非线性关系的其中一个可能性，用参数来表述线性关系。

[7] 统计软件 Stata 的 R^2_L 就是伪 R^2（Stata，1999）。旧版的 SAS PROC LOGISTIC [SAS（SUGI）PROC LOGIST（Harrell，1986）] 根据模型参数的数量调整了 R^2_L，这种类似线性回归的调整后的 R^2，写成 R^2_{LA}，表示它与 R^2_L 有联系并区别于其他 R^2 量表。$R^2_{LA} = (G_M - 2k)/D_0$，其中，$k$ 是模型中自变量的数量。如果 $G_M < 2k$，尤其当 $G_M = 0$ 时，用 R^2_{LA} 可能出现负数的已解释方差估计。

[8] 应当注意，在其他情况下，它可能是不恰当的。例如，在比例风险模型中，R^2_L 对审查比 R^2_N 敏感（Schemper，1990，1992）。

[9] 选用 ϕ_p，如 ϕ，它基于每格的观测和预期值的比较（而不是列或栏如 λ_p 和 τ_p），如果表格的边缘值一致（列和栏的总和，大的列总和对应大的栏总和），ϕ_p 接近 ϕ 的数值，那么 ϕ_p 与 ϕ 符号就一样（因为它们分子相同）。

[10] 调整 ϕ_p，可以把误差的最小值 $|(a+b)-(a+c)| = |b-c|$，加上没有模型的误差预期值。系数会成：(a) 仍可解释错误的比例改变（因为调整了预期误差的计算）；(b) 如果该模型是不准确的话，仍会出现负数。极度差劣的模型，已修订的指数的最大值小于 1，即使分类准确度极高，但 ϕ_p 的增量却非常小。基于与 ϕ 相似，我们把它写成 ϕ'_p。注意，不能算成 $\phi_p/\max\phi_p$；这样做会破坏量表的错误比例改变的解释，而当 ϕ_p 的最大值是 0，量表会变成无定义 0。

[11] 对于两尾检验，原假设有无预测模型的误差没有任何区别。另一种假说是有预测模型的误差比例不等于没有预测模型的误差。一尾检验指模型会提高预测因变量的精确度，原假设是预测模型的误差比例不小于没有模型的误差比例。另一种假说是预测模型的误差比例少于没有模型的误差比例。预测模型是否能提高个案的分类能力，一尾检验较为适当，负 λ_p 将导致负 d 值和不拒绝原假设。

[12] Copas 和 Loeber（1990）指出，在这种情况下，1 数值会被误解成完美的预测。两个问题：第一，在

这种情形下，我们应如何解释 RIOC 的数值？第二，什么 RIOC 数值才算是完美预测？如果有任何不理想的量表变化和更好替代品，就不要用 RIOC。

[13] 二分因变量 logistic 回归的 R^2 并不经常比线性回归的 R^2 高。在有关小偷的全国青年调查研究中，线性回归 R^2 和 logistic 回归 R^2 分别是 0.255 和 0.253。

[14] 这有时被称为一类错误或假阴性（未能检测到存在的关系），相对于一类错误或假阳性（总结有关系，但实际没有）。

[15] 后面逐步的程序在文中已讨论。SPSS NOMREG 和 PLUM 没有逐步的程序，但它的输出可包括似然统计。

[16] 如果有的话，他们可表明非西班牙欧裔美国人使用大麻率最低，其次是非裔美国人，其他人使用大麻最高。这令人怀疑，但是，声明这一关系没有统计显著性，可能只是随机样本的误差。

[17] 从数学的角度看，省略掉的类别是多余的或不是研究的兴趣。但理论检验和应用研究都认为提供系数的全部资料及其统计意义是比较合理的，而不是纸笔的计算。

[18] 霍斯默和莱默苏（1989）区分了基于个案的残差分析和共变形式（样本自变量的数值组合）的残差分析。当共变形式数目等于个案数目，或非常接近，每个个案的残差必须分开分析。这是 SAS PROC LOGISTIC 和 SPSS LOGISTIC REGRESSION 所隐含的方法，但 SPSS NOMREG 用共变形式的集合数据计算正确的预测、残差和拟合适度（Norusis, 1999）。当个案数量远远超过共变模式，或共变量模式具有多于 5 个个案的话，霍斯默和莱默苏建议透过共变模式集合数据，避免低估杠杆统计 h_j。

[19] 标准正态分布（平均值为 0 和标准偏差为 1）95%的个案的标准分数（或标准残差）应在 −2 与 2 之间，99%在 −3 与 3 之间。标准或偏离残差大于 2 或 3 不一定表示模型有问题，但如果约 5% 的样品超出 −2 至 +2 的范围，1%在 −2.5 至 +2.5 内。数值远超出这一范围，通常表示某些个案与模型拟合不好，因此，需要进一步分析或修改模型。

[20] 如福克斯（Fox, 1991）指出，在线性回归中，影响力＝杠杆×差异，其中"差异"是指 Y 对预测值的离群值。logistic 回归相对于线性回归，拟合概率越亲近 0（小于 1）或 1（大于 0.9），它的杠杆停止增加并迅速转向 0（Hosmer & Lemeshow, 1989：153—154）。

[21] 请更清楚地了解过度分散和过度集中的讨论，可参见麦卡拉和内尔德（1989：124—128）。

[22] SAS CATMOD 不会直接计出 D_0 和 G_M。这两者的似然率 χ^2 V 统计是根据列联表中各格的比较，而不是分类概率。但是这两者是相关的，它可以从 SAS PROC CATMOD 的统计推出适当的 logistic 回归统计，步骤如下：

(1) 计算 $D_0 = \sum(n_{Y=h})\ln[P(Y=h)] = \sum(n_{Y=h})\ln(n_{Y=h}/N)$，$n_{Y=h}$ 是 $Y=1$ 的个案数目，N 是总样本数量，以及所有 $Y=h$ 的可能值的总和。

(2) 检查模型的迭代历史。在最后迭代中最大似然分析表的"−2 对数似然"接近（但不完全）等于 D_M。

(3) 计算 $G_M = D_0 - D_M$；计算 $R_L^2 = G_M/D_0$。

如果二分因变量分析沿用这些程序，那么结果的数据大约等于 SAS PROC LOGISTIC 的 G_M 和 R_L^2（SAS PROC LOGISTIC 用再加权最小二乘算法去计算模型参数；PROC CATMOD 使用加权最小二乘或最大似然，这根据模型而定）。虽然这样发生比有点怪，但是如果 CATMOD 的似然率没修改的话，当因变量是二分时，其结果就会与 SAS PROC LOGISTIC 不同。

[23] 要想有可信的估计，就需要一个大的样本（Hu, Bentler & Kano, 1992）。

参考文献

AGRESTI, A. (1990). *Categorical data analysis*. New York: Wiley.

AGRESTI, A., and FINLAY, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

ALDRICH, J. H., and NELSON, F. D. (1984). *Linearprobability, logit, andprobit models*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—045. Beverly Hills, CA: Sage.

ALLISON, P. D. (1999). *Logistic regression using the SAS system*. Cary, NC: SAS Institute.

BEGG, C. B., and GREY, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71*, 11—18.

BENDEL, R. B., and AFIFI, A. A. (1977). Comparison of stopping rules in forward regression. *Journal of the American Statistical Association*, *72*, 46—53.

BERRY, W. D. (1993). *Understanding regression assumptions*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—092. Newbury Park, CA: Sage.

BERRY, W. D., and FELDMAN, S. (1985). *Multiple regression in practice*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—050. Beverly Hills, CA: Sage.

BOHRNSTEDT, G. W., and KNOKE, D. (1994). *Statistics for social data analysis* (3rd ed.). Itasca, IL: F. E. Peacock.

BOLLEN, K. A. (1989). *Structural equation models with latent variables*. New York: Wiley.

BULMER, M. G. (1979). *Principles of statistics*. New York: Dover.

CLOGG, C. C., and SHIHADEH, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

COPAS, J. B., and LOEBER, R. (1990). Relative improvement over chance (RIOC) for 2×2 tables. *British Journal of Mathematical and Statistical Psychology*, *43*, 293—307.

COSTNER, H. L. (1965). Criteria for measures of association. *American Sociological Review 30*, 341—353.

COX, D. R., and SNELL, E. J. (1989). *The analysis of binary data* (2nd ed.) London: i Chapman and Hall.

CRAGG, J. G., and UHLER, R. (1970). The demand for automobiles. *Canadian Journal of Economics 3*, 386—406.

DeMARIS, A. (1992). *Logit modeling*. Sage University Paper Series on Quantitative I Applications in the Social Sciences, 07—086. Newbury Park, CA: Sage.

ELIASON, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—096. Newbury Park, CA: Sage.

ELLIOTT, D. S, HUIZINGA, D., and AGETON, S. S. (1985). *Explaining delinquency and drug use*. Beverly Hills, CA: Sage.

ELNLIeOwT YTo, rDk:. SSp. , r iHngUeIrZ-VINerGlaAg. , D. , and MENARD, S. (1989). *Multiple problem youth.* II

FARRINGTON, D. P. , and LOEBER, R. (1989). Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2×2 tables. *Journal of Quantitative Criminology*, *5*, 201—213.

FOX, J. (1991). *Regression diagnostics.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—079. Newbury Park, CA: Sage.

HAGLE, T M. , and MITCHELL, G. E. , II (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, *36*, 762—784.

HARDY, M. (1993). *Regression with dummy variables.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—093. Newbury Park, CA: Sage.

HARRELL, F. E. , Jr. (1986). The LOGIST procedure. In SAS Institute, Inc. (Ed.), *SUGI supplemental library user's guide*, (Version 5, pp. 269—293). Cary, NC: SAS Institute.

HOSMER, D. W. , and LEMESHOW, S. (1989). *Applied logistic regression.* New York: Wiley.

HU, L. , BENTLER, P. M. , AND KANO, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351—362.

HUTCHESON, G. , and SOFRONIOU, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* Thousand Oaks, CA: Sage.

JORESKOG, K. G. , and SORBOM, D. (1988). *PRELIS: A program for multivariate data screening and data summarization* (2nd ed.). Chicago, IL: Scientific Software International.

JORESKOG, K. G. , and SORBOM, D. (1993). *LZSREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International. KLECKA, W. R. (1980). *Discriminant analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—019. Beverly Hills, CA: Sage.

KNOKE, D. , and BURKE, P. J. (1980). *Log-linear models.* Sage University Paper.

Series on Quantitative Applications in the Social Sciences, 07—020. Beverly Hills, CA: Sage.

LANDWEHR, J. M. , PREGIBON, D. , and SHOEMAKER, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, *79*, 61—71.

LEWIS-BECK, M. S. (1980). *Applied regression: An introduction.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—022. Beverly Hills, CA: Sage.

LOEBER, R. , and DISHION, T. (1983). Early predictors of male delinquency: A review. *Psycholocal Bulletin*, *94*, 68—99.

LONG, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

MADDALA, G. S. (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge, UK: Cambridge University Press.

MAGEE, L. (1990). *R2* measures based on Wold and likelihood ratio joint significance tests. *The American Statistician*, *44*, 250—253.

McCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models* (2nd ed.). London:

Chapman and Hall.

McFADDEN, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, *3*, 303—328.

McKELVEY, R., and ZAVOINA, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Socioloa*, *4*, 103—120.

MENARD, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*, 17—24.

MIECZKOWSKI, T(1990). The accuracy of self-reported drug use: An evaluation and analysis of new data, In R. Weisheit (Ed.), *Drugs, crime, and the criminal justice system* (pp. 275—302). Cincinnati: Anderson.

NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691—692.

NORUSIS, M. J. (1999). *SPSS regression models 10.0*. Chicago: SPSS, Inc.

OHLIN, L. E., and DUNCAN, 0. D. (1949). The efficiency of prediction in criminology. *American Journal of Sociology*, *54*, 441—451.

SAS(1989). *SASISTAT user's guide* (Version 6, 4th ed. Vols. 1 and 2). Cary, NC: SAS Institute.

SAS(1995). *Logistic regression examples using the SAS system*. Cary, NC: SAS Institute.

SCHAEFER, R. L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, *25*, 75—91.

SCHEMPER, M. (1990). The explained variation in proportional hazards regression. *Biometrika*, *77*, 216—218.

SCHEMPER, M. (1992). Further results on the explained variation in proportional hazards regression. *Biometrika*, *79*, 202—204.

SCHROEDER, L. D., SJOQUIST, D. L., and STEPHAN, P. E. (1986). *Understanding regression analysis: An introductory guide*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07—057. Beverly Hills, CA: Sage.

SIMONOFF, J. S. (1998). Logistic regression, categorical predictors, and goodness of fit: It depends on who you ask. *The American Statistician*, *52*, 10—14.

SODERSTROM, I., and LEITNER, D. (1997). The effects of base rate, selection ratio, sample size, and reliability of predictors on predictive efficiency indices associated with logistic regression models. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago.

SPSS(1991). *SPSS statistical algolithms* (2nd ed.). Chicago: SPSS, Inc.

SPSS(1999a). *SPSS advanced models 10.0*. Chicago: SPSS, Inc.

SPSS(1999b). *SPSS base 10.0 applications guide*. Chicago: SPSS, Inc.

STATA(1999). *Stata reference manual* (Release 6, Vol. 2). College Station, TX: Stata Press.

STUDENMUND, A. H., and CASSIDY, H. J. (1987). *Using econometrics: A practical guide*. Boston: Little, Brown.

VEALL, M. R., and ZIMMERMAN, K. F. (1996). Pseudo-R2 measures for come common limited dependent variable models. *Journal of Economic Surveys*, *10*, 241—260.

WIGGINS, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

WOFFORD, S., ELLIOTT, D. S., and MENARD, S. (1994). Continuities in marital violence. *Journal of Family Violence*, *9*, 195—225.

译名对照表

Analysis of Variance	方差分析
Backward Elimination	后淘汰法
Base Rate	基准比率
Bivariate Regression	二元回归
Canonical Correlation	典型相关
Casewise Approach	个案为主
Coefficient of Determinant, R^2	决定系数
Collinearity	共线性
Conditional Mean	条件平均数
Conditional Probabilities	条件概率
Confidence Interval	置信区间
Contingency Table Approach	列联表方法
Continuous Scale	连续量表
Converge	收敛
Covariate Pattern	共变规则
Covariates	协变量
Criterion Variable	基准变量
Crosstabulation	交叉列表分析
Degree of Freedom	自由度
Dependent Variable	因变量
Deviance	偏差
Deviation Coding	离差编码
Deviation Statistics	偏离统计
Dichotomous Predictors	二分预测变量
Discrimination Analysis	判别分析
Dummy Variable	虚拟变量
Endogenous Variable	内生变量
Exogenous Variable	外生变量
Explained Variance	已解释方差
Exponentiated Coefficients	指数化系数
Exponentiation	指数方式
False Negative	假阴性
Forward Inclusion	向前包括法
Full Model	全模型
Goodness of fit	拟合优度

Grouped data	分组数据
Heterogeneous	异质的
Homogeneous	同质的
Homoscedasticity	同方差
Hypothesis Testing	假设检验
Independent variable	自变量
Indicator Coding	指针编码
Interaction	交互项
Intercept	截距
Interval Scale	定距量表
Iteration	迭代
Iterative Process	迭代过程
Juvenile Gangs	童党
Leverage	杠杆
Leverage Statistics	杠杆统计
Likelihood Ratio Test	似然比检验
Linear Probability Model	线性概率模型
Linear Regression Model	线性回归模型
Logistic Regression logistic	回归
log-likelihood	对数似然
Marginal Distribution	边际分布
Maximum likelihood	最大似然法
Maximum Likelihood Estimation	最大似然估计
Measurement	测量
Measurement Assumption	测量假设
Missing Data	缺失值
Model Chi-square	模型卡方
Model Fitting Information table	模型拟合数据表
Model if Term Removal table	删除变量后的模型表
Model Summary	模型概要表
Multiple Regression	多元回归
Natural Logarithm	自然对数
Nomial Regression	名义回归
Nominal Variable	名义变量
Nonlinearity	非线性
Normality	正态性

Null Hypothesis	原假设
Odds ratio	优比
Ominibus Tests of Model Coefficients	模型系数综合检验表
Ordinal Varible	定序变量
Ordinal Logistic Regression	定序
Ordinary Least Square, OLS	普通最小二乘法
Orthogonal	直角正交
Outcome Variable	结果变量
Outlier	奇异值
Output	输出
Parameter Estimates	参数估计
Partial Slope Coefficient	偏斜率系数
Polynomial	多项
Polytomous Logit Universal Model	多分类
Logit	通用模型
Polytomous Nominal	多分类
Polytomous Nominal Logistic Regression	多分类名义
logistic	回归
Population parameters	总体参数
Predicted probability	预测概率
Predictive efficiency	预测效率
Predictor Variable	预测变量
Profile Contrast	形象对比
Proportional Change in Error	误差改变比率
Proportional reduction in error	误差减少的比例
Pseudo-R^2	伪 R^2
Ratio	比例
Ratio Scale	定比量表
Regression Coefficient	回归系数
Relative Improvement over Chance, RIOC	相对机会提高率
Repeated Contrast	反复对比
Residuals	残差
Reversed Helmert	反向赫尔默特
Ridge Regression	岭回归
Risk ratios	风险比
SAS Display Manager	SAS 显示管理系统

Saturated Model	饱和模型
Simple Coding	简单编码
Slope	斜率
Specification	确定性
Specification Error	设定误差
Standard Error of Estimate	估计标准误差
Stepwise	逐步法
Studentized Residual	学生氏残差
Sum of the Squared Errors of Prediction	预测平方误差总和
Sums of Squared Errors	误差平方和
Suppressor Effect	抑制效应
t ratio	t 比率
Total Sum of Squares，SST	总平方和
Unexplained Variance	未解释方差
Variables in the Equation	方程内生变量

定序因变量的 logistic 回归模型

作 者 简 介

安·A. 奥康奈尔(Ann A. O'Connell)

美国康涅狄格斯托斯大学教育心理学副教授,同时也是测量、评估和评价研究生项目的协调者。教授研究生统计学和教育研究方法,包括多层次模型、logistic 回归、抽样和调查研究方法、多变量分析等,并关注这些方法和其他高级统计/研究方法的应用。曾在 *Women and Health*, *Evaluation and the Health Professions*, *Measurement and Research in Counseling and Development*, *Morbidity and Mortality Weekly Report*,以及 *Journal of Educational Research* 等期刊上发表过文章。

译 者 简 介

赵亮员

北京大学社会学系 2007 级博士研究生。

序

在过去60年中，logit模型已经成为社会科学中最流行的统计方法。自从伦西斯·利克特(Rensis Likert)1932年出版《态度测量的方法》(*A Technique for the Measurement of Attitudes*)以来，人们对人类态度的测量方法一直都在变化着。实际上，今天的任何社会调查都将利克特类型的量表作为一种主要的方法。一个典型的利克特量表有五个类别(例如：非常不同意、同意、不知道、同意、非常同意)，用以测量个体对问题的响应，尽管它在别的地方也许还有3到7个或更多的响应类别。如果我们将五个类别编码为从1到5，那么我们可以用线性回归模型来估计利克特量表，这也是早些时候分析这样的数据时所采用的方法。然而，这种方法也存在诸多明显的问题。首先，经典线性回归假设连续因变量是距离相等的定序因变量类别(ordered response categories)。一个利克特类型的量表，或者任何其他的定序量表，虽然是定序的，但并不一定在类别间等距。其次，也许是更重要的，这样的量表所得出的分布不是经典线性回归所假定的数据呈现的正态分布。

为了分析定序数据的这种属性，也有其他的方法，经常使用的是列联表和对数线性模型的形式，具体可以参见：《定序数据分析》(*Analysis of Ordinal Data*)，作者海德普兰德、莱恩和罗森塔尔(Hiderbrand, Laing, Rosenthal)；《对数线性模型》(*Log-linear Models*)，作者诺克和伯克(Knoke, Burke)；《潜变量的对数线性模型》(*Loglinear Models With Latent Variables*)，作者哈根纳斯(Hagenaars)；《定序对数线性模型》(*Ordinal Log-Linear Models*)，作者石井坤茨(Ishii-Kuntz)；《列联表分析中的比数比》(*Odds Ratios in the Analysis of Contingency Tables*)，作者鲁达斯(Rudas)。然而，这些方法都不是在回归的框架下，而回归方法正是社会科学中广为人知且被广泛应用的量化方法。

安·A.奥康奈尔的这本书填补了空白。即使有的著作也将定序因变量变量在logit模型中加以处理，但本书则通过呈现三种形式的因变量来完全集中关注该类logit模型，这些因变量抓住了响应的定序本质。本书以展示来自儿童早期追踪研究的实例开始，其中的主要因变量虽然不是利克特量表，但其对早期读写和算术熟练度的测量还是定序的。作者进而回顾了logistic回归，并接着呈现三个核心主题章节：累积或比例比数模型

(cumulative or proportional odds model)；连续比例模型(continuation ratio model)以及相邻类别模型(adjacent categories model)。与此同时，书中给出了 SAS® 和 SPSS® 的案例。尽管比例比数模型也许是三种方法中最常用的，但在结论部分，读者可以领会到关于替代方法尤其是什么时候使用替代方法的提示。

——廖福挺(Tim Futing Liao)

第 1 章 | 概述

对于很多在教育和社会科学中的因变量，定序尺度提供了一种将可能的结果区分开来的简单且便捷的方法，而这些可能结果最好被看成是等级定序的(rank-orderd)。定序数据的主要特点，是指定给所测量变量的相继类别的数字代表的是量级上的差异，或者是"大于"或者是"小于"的性质(Stevens，1946，1951)。一些定序数据的例子包括：对文章的赋值指示以及那些基于改善的质量水平而对响应进行打分的算法问题的解决方案(例如，0 = 差，1 = 可以接受，2 = 优秀)。对比之下，名义水平的数据发生在数值用于通过认定类别之间简单的质的差别来测量一个变量的情况(例如，1 = 男性，或者 2 = 女性；所上学校的地理位置描述：1 = 乡村，2 = 城市，3 = 郊区，等等)；名义变量没有定序数据的方向特征。另一方面，在定距或者定比水平尺度上测量的变量的确使用了表明定序数据"大于"或者"小于"属性的尺度值，但还保持了在尺度间相邻值距离均等或者间隔均等长度的属性。以摄氏尺度测量的温度是我们熟知的定距水平变量的例子。然而，定距水平的变量的零点相对主观，而非绝对。拥有所有定距尺度属性但也有真实零点的变量，被称为定比水平。例如，对一项任务的反应时间、重量以及距离是熟悉的定比水平的变量。[1]

定序类别在将数字指派为代表贡献、建构或者行为的相继类别且有方向意义上的差异的研究情境下很常见。纳普(Knapp，1999)使用定序评级法，通过划分轻微的(1)、适中的(2)、严重的(3)尺度类别来评估疾病的严重程度。在纳普的研究中，归于疾病严重度类别的数字代表了逐渐增加的严重度，在这个意义上"适中"比"轻微"更严重，而"严重"比"适中"更严重。赋予"严重"案例的数值并不是说"严重"是"轻微"的三倍严重程度，而是说"严重"的疾病严重度比

"轻微"类别的疾病严重度要大,也比"适中"类别的那些要大。

用来代表进度上更严重的类别的数字选择很方便地保留了定义类别本身的"大于"或者"小于"的隐含属性。数字是对研究的属性进行建模,例如疾病的严重程度,并被用来保留类别的传递性;如果值(3)代表着比值(2)严重得多的状态,而值(2)又代表着比值(1)严重得多的状态,那么传递性的特征意味着值(3)代表的状态比值(1)代表的状态要严重得多(Cliff & Keats, 2003; Krantz, Luce, Suppes & Tversky, 1971)。

对定序尺度变量的测量是类似的。定序尺度在咨询服务和精神疗法中已经用于主观的概率或者可能性的判断中(例如,对从 1 = "非常不可能"到 5 = "非常可能"进行评分)(Ness, 1995)。一个客户在经过治疗后的临床表现可以归类为:恶化(1)、没有变化(2)、改善(3)(Grissom, 1994)。健康研究者经常使用几个相继的水平来描述健康相关行为的"改变阶段",例如,戒烟、使用避孕套、运动行为、努力减肥(Hedeker & Mermelstein, 1998; Plotnikoff, Blanchard, Hotz & Rhodes, 2001; Prochaska & DiClemente, 1983, 1986; Prochaska, DiClemente & Norcross, 1992)。在改变阶段模型中,朝着行为改变的意向或者行为通常通过以下阶段来测量:(1)意图前期(precontemplation);(2)意图期(contemplation);(3)准备期(preparation);(4)行动期(action)和(5)维持期(maintenance)。教师对在他们的教室里所实施的教育改革的关注体验也通过定序尺度来测量,它代表了关注的焦点从 1(自身)到 7(他人)(Hall & Hordes, 1984; van den Berg, Sleegers, Geijsel & Vandeberghe, 2000)。在童年早期教育中,早期读写能力的层次体的掌握指标会影响青少年儿童的读写熟练度,本质上可以通过定序来归类:(1)语音语韵觉识(phonemic awareness);(2)发音(phonics);(3)流利阅读(fluency);(4)词汇(vocabulary);(5)文本理解(text comprehensive)(早期阅读能力改善中心,Center for the Improvement of Early Reading Achievement, CIERA, 2000)。

尽管定序结果可以是简单且有意义的,但对它们进行理想的统计处理对很多实证研究者而言仍然是一项挑战(Cliff, 1996a; Clogg & Shihadeh, 1994; Ishii-Kuntz, 1994)。历史上,研究者对定序结果的分析依赖于两种非常不同的

方法。一些研究者选择使用定序结果的参数模型，例如通过将定序结果作为至少定距水平的变量的多元线性回归，假设这些方法的稳健性克服了任何可能的解释问题。其他的研究者选择将定序变量严格地当做类别变量对待，并运用对数线性或非参数方法来理解数据。尽管两种策略在研究问题的基础上都能够说明问题，但它们都不是发展出定序结果的解释模型的最佳方法（Agresti，1989；Cliff，1996a；Clogg & Shihadeh，1994；O'Connell，2000），尤其是当分析的焦点是在这些定序分数的区分方面时。

本书的目的

本书的目的是让实证研究者熟悉，尤其是让那些在教育和社会科学及行为科学领域的研究者们熟悉，忠实于结果测量实际水平的定序因变量的替代分析方法。本书讨论的例子是定序回归模型，它们是对二分响应数据的 logistic 模型的扩展。logistic 回归方法是在流行病学、医学和相关领域内稳固发展起来的，实际上，最近关于定序回归技术应用和发展的文献也大都出现在公共卫生研究团体内。很多这类统计或者比较研究的结果会在这里提到。教育和社会科学家也许通常不关注类似于流行病学或者医学研究者研究的变量，但是困扰这两个领域的模型的适用性问题，其大部分都可以从更广泛的统计文献中应用不同方法解决统计难题中了解到。

本书将呈现分析定序结果数据的三种方法，并通过例子进行描述。这些方法包括比例或累积比数模型（Proportional or Cumulative Odds model，以下简称 CO）（Agresti，1996；Armstrong & Sloan，1989；Long，1997；McCullagh，1980），连续比例模型（Continuation Ratio model，以下简称 CR）（Armstrong & Sloan，1989；D. R. Cox，1972；Greenland，1994）以及相邻类别模型（Adjacent Categories model，以下简称 AC）（Agresti，1989；Goodman，1983）。此外，作者呈现了偏比例比数的例子（Peterson & Harrell，1990），并讨论了偏比例风险或者未约束的连续比数模型（Bender & Benner，2000；Cole & Ananth，2001）作为比例比数或者连续比例模型的假设被违背情况下的替代分析。

定序 logit 模型可以看作二分结果的 logistic 回归的扩展，相应的，这些模型密切遵循 logistic 和普通最小二乘回归分析的方法和模型建构策略。作者决定关注定序结果的 logit 模型，是因为从这些模型中得到的概率和比数的解释较为直观。例如，这里呈现的方法的替代，包括安德森（Anderson，1984）的定型模型（stereotype model）、probit 回归模型，以及定序结果变量的结构方程模型的多分格相关（polychoric correlation）的应用。这些和其他定序数据分析的策略在 Huynh（2002），Borooah（2002），Ishii-Kuntz（1994），Liao（1994），Menard（1995）以及 Jöreskog 和 Sörbom（1996）的研究中都可以看到；大体上，关于定序变量的处理可以参考：Long（1997），Clogg & Shihadeh（1994）和 Agresti（1989，1996）的论著。

累积比数模型是最常使用的定序回归模型，尽管这里考察的所有模型很多研究者仍然不太熟悉，尤其是在教育科学领域。作者所回顾的各个模型能够解决存在于定序结果研究中的独特问题，且这些问题通过将数据作为定距/定比或者严格的分类数据时并不能得到满意的结果。

软件和句法

这里使用 SAS® 和 SPSS® 统计软件来描述例子。在各个软件里，作者使用 SAS PROC LOGISTIC（升序或者降序选项），SAS PROC GENMOD，SAS PROC CATMOD，SPSS LOGISTIC REGRESSION，以及 SPSS PLUM 来运行不同的例子。本书的附录包括各个分析的句法，这些句法和数据都可以在作者的网站上找到：http://faculty.education.uconn.edu/epsy/aconnell/index.htm[①]。全书在必要的时候将会注明各种统计软件的局限、类似之处和差异。这里呈现的所有分析都假设儿童之间的独立。在本书最后一章，作者简要讨论了多层次数据的定序因变量的变量处理方法，这是一个迅速增长的领域，它由单层次数据的比例比数模型和一般多层次模型研究逻辑扩展而来。

① 至翻译完成时该链接已失效，需要数据的读者可以向译者或者出版社电邮索取。

作者在本书中主要使用 SAS 和 SPSS 来描述定序 logit 模型的概念和程序。另一个分类数据分析的全面的统计软件包，也包括定序数据分析的优秀模块是 Stata(Long & Freese, 2003)。Stata 也包括能够促进对这里呈现的模型的进一步理解的做图能力。不论选择何种统计软件，本书中的模型描述都是适用的。

本书结构

第 2 章描述了这里呈现的分析所使用的数据集。第 3 章包括对 logistic 回归分析的概要回顾，澄清了对于理解 logit 类型的定序回归模型而言相当重要的术语，这些术语包括：比数(odds)、比数比(odds ratios)、logits，以及模型拟合(model fit)。基于它们与 logistic 回归模型概念上的类似性，第 4 到第 6 章依次描述了三个定序模型(CO, CR 和 AC)。对于这里呈现的每一个定序模型，都会解释模型和变量作用，并讨论模型拟合的评估和预测效率。第 4 章提供了与一般最小二乘多元回归之间的比较。最后，第 7 章回顾并概述了这里所研究的分析，并讨论了这些模型的一些扩展。各个分析的一些电脑输出也将包括在内。

本书中的数据都来自于儿童早期追踪研究—幼儿园同期群(Early Childhood Longitudinal Study-Kindergarten Cohort, ECLS-K)，该研究跟踪了代表全国样本的幼儿园儿童的早期阅读和算术的进步情况，直到一年级结束(三年级的数据在 2004 年 3 月开放)。这里分析的数据是进入一年级时候的。ECLS-K 数据是由美国国家教育统计中心(NCES)执行的，也评估了学生早期读写、算术以及综合知识能力的熟练度，作为“石阶”(stepping-stones)系列的一部分，能够反映形成今后学习基础的定序能力(West, Denton & Germino-Hausken, 2000)。所有的可供公众使用的一年级数据库都可以从 NCES 处得到[2]。这里描述的例子推导仅是为了阐明定序回归模型的技术和方法使用的目的，并不是为了提供关于一年级儿童早期阅读成就的完整画面。有关儿童早期阅读的影响因素的更多信息，可以参见例如斯诺、伯恩斯及格里芬的著作(Snow, Burns & Griffin, 1998)。

第 2 章 | 背景介绍:儿童早期追踪研究

儿童早期追踪研究概述

儿童早期追踪研究提供了一个一年级儿童的完整画面,包括他们的幼儿园经历、早期家庭经历及其老师和学校。ECLS-K 数据考察了识字、阅读以及算术能力。它包括一套对项目反应理论(Item-Response Theory, IRT)尺度的认知评估系列,是在全国层面样本学校的近两万名儿童中进行的。除了常模参照的连续 IRT 测量,ECLS-K 还包括标准参照学生的识字和算术熟练度,它通过学生对一系列问题群(包括 5 组、每组 4 道题目)的回答,来反映作为后续阅读和数学学习的基础。所得的结果分数可以用来在个体层次上对学生进行诊断,并找出针对个体的干预方向,在群体层次上,也可以用来建议针对处于不同熟练度层次的学生的可能干预方式。本书中所讨论的分析将集中在表示读写熟练度的标准参照分数上。

ECLS-K 评估工具中所呈现的早期读写熟练度的类别与已被界定的阅读熟练度的基石(building blocks)一致:语音语韵觉识理解代表口语发音的字母、发音(理解字母组合的发音)、流利阅读、词汇、文本理解(CIERA, 2001)。隐含在读写发展背后的技能是有层级且互相依赖的;稍后的技能在之前的技能得到发展之前是不大可能实现的。表 2.1 描述了 ECLS-K 所使用的熟练度类别。

充分回答代表各个类别的题项群的能力被假设为服从古特曼模型(Guttman, 1954; NCES, 2000, 2002);也就是说,掌握了一个层次即默认为掌握了所有之前的层次。在 ECLS-K 评估中,样本中的每个儿童都在各个题项群上得到一个通过或者失败的分数,这些题项群代表了熟练度水平(从 1 到 5),直到该儿

童未能通过某题项群中 4 个问题中的 3 个。[3]对一个水平的掌握表明对所有之前水平的掌握；一旦儿童不能成功通过某题项群，则测试中止。[4]因此，在 ECLS-K 数据库中阅读熟练度有 5 个二分变量(C3RRPRF1 到 C3RRPRF5)。例如，如果一个儿童填过了读写水平 1 中 4 个问题中的 3 个和读写水平 2 中 4 个问题中的 3 个，则该儿童的 C3RRPRF1 和 C3RRPRF2 将会得到为 1 的值。如果还是该儿童，未能通过下一题项群(读写水平 3)中 4 个问题中的 3 个，则 C3RRPRF3 则为 0，后续的水平也为 0。对于这里展示的分析，五个二分熟练度值的系列用来产生一个反映了对定序尺度内容的掌握的变量。在重新编码以获得一个定序变量后，以上假定的学生将会得到一个为 2 的值，作为他或者她的熟练度得分，代表了对材料的掌握可达 2 级(包括 2 级)。以这种方式，在 ECLS-K 样本中的每个儿童都得到一个单独的变量(profread)以评估读写熟练度，该变量有 6 个可能的结果类别(从 0 级到 5 级)。该定序尺度得分为 0 时，表示该儿童没有取得代表熟练度 1 级的题项群的掌握程度。[5]

表 2.1　ECLS-K 对早期读写测量的熟练度类别

熟练度类别	描　　述
0	没有通过 1 级
1	能够认清大小写字母
2	能够将字母与单词开头的发音联系起来
3	能够将字母与单词末尾的发音联系起来
4	能够识别常见单词
5	能够阅读文本中的单词

资料来源：美国国家教育统计中心(NCES，2002)。

定序结果的实际重要性

定序熟练度分数可以为研究者和教育学家揭示儿童在继续他们的小学求学过程中，在通往完全掌握读写能力的路上还有多远。将定序熟练度分数而非连续 IRT 尺度分数作为分析的兴趣变量，强调了熟练度评估在识别和选择学生以进行早期干预项目中所扮演的角色。这些分析可以给出一些在需要个性化

的干预以满足特定学生需求的具体领域的建议。定序熟练度结果实际上是通常的定序变量，就它们把干预导向特定层次的熟练度方面而言，有着巨大的实用价值。对于班级中的老师或者阅读专家，熟练度分数也许比知道儿童的认知评估的 IRT 尺度分数为 55 分更有价值，也更具可解释性。根据班级或者学校的实际制定政策进行的干预，将会与成功获得读写技能的石阶(stepping-stone)相关联，对于个体的学生也比那些急于改善全局认知测试分数(基于教室、学校或者区的层次)的尝试更有效。

模型中的变量

这里展示的分析中被选为预测值的变量与幼童的早期阅读能力相关联。ECLS-幼儿园同期群的初始数据概要表明：一些儿童比其他儿童在学习方面表现出更多的准备(preparedness)和"就位"(readiness)，将他们置于重要的学校低年级阶段同龄人的前一步(NCES, 2000)。ECLS-K 研究已经展示了那些有着某些特征的入园儿童(生活在单亲家庭、生活在接受福利金或者食物券的家庭、母亲的教育水平低于高中，或者父母的主要语言不是英语)倾向于处于低的阅读技能风险中(Zill & West, 2001)。与家庭相关的入园前的经历(如有父母为其阅读)，进入学前班或者得到日间看护，以及个人特征(如性别)等，都与儿童在阅读中的初始熟练度相关，也与幼儿园及之后年级中儿童在技能和能力上的潜在增长相关。例如，通常入园的女孩比男孩具备稍好一些的阅读能力。基于儿童早期阅读成功失败的预测值对于理解个体儿童如何处于阅读困难的问题有帮助；从政策和实践的角度看，很明显需要教师、学校当局、父母以及其他利益相关者意识到这些与早期熟练度相关的个体因素，从而发展出并支持能够促进所有儿童取得相对于他们一年级和入园时的技能更大成就的课程设置和教学实践。

六个熟练度类别间解释变量的描述性统计结果在表 2.2 中给出。它们包括 *gender*(性别)，以男性的%呈现(0 = 女性，1 = 男性)，*risknum*(家庭危险因素个数，从 0 到 4，基于父母特征，包括住在单亲住户、住在接受福利金或食物券的家庭、母亲的教育低于高中，或者父母的主要语言不是英语)，*famrisk*(表明

家庭危险是否出现的二分变量，0 为无，1 为有或者说 *risknum* 大于等于 1)，*plreadbo*(在入园前经常有父母为其阅读书籍，从 1 到 4 打分，1 = 从不，4 = 每天)，*noreadbo*(二分变量，0 表示父母为其阅读的频率从每周三次或以上到每天都会，1 表示每周阅读不到 1 次或者 2 次)，*halfdayK*(儿童上的是半天还是全天的幼儿园，0 = 否，即全天幼儿园，1 = 是，即半天幼儿园)，*center*(入园前儿童是否上过基于中心的日间看护，0 = 否，1 = 是)，*minority*(少数族裔，0 = 白人/高加索背景；1 = 少数族裔或任何其他的背景)，*wksesl*(入园之前的家庭社会经济地位，连续尺度分数，均值为 0)，以及 *plageent*(入园时儿童的年龄月数)。另外一个变量，*public*(儿童所上的学校类型，0 = 私立，1 = 公立)因为设计方面的担心，包括在描述中但没有包括在模型中。

表 2.2　进入一年级时的描述统计，N = 3365

	Reading Proficiency Level(profread)						
	0 (n = 67)	1 (n = 278)	2 (n = 594)	3 (n = 1482)	4 (n = 587)	5 (n = 357)	Total (N = 3365)
%*profread*	2.0%	8.3%	17.7%	44.0%	17.4%	10.6%	100%
%*male*	71.6%	58.6%	53.9%	49.6%	43.6%	42.3%	49.7%
risknum							
M	0.97	0.77	0.65	0.44	0.32	0.25	0.47
(*SD*)	(1.04)	(0.88)	(0.88)	(0.71)	(0.61)	(0.53)	(0.75)
%*famrisk*	58.2%	52.9%	43.8%	32.5%	25.9%	20.7%	34.3%
%*noreadbo*	38.8%	27.0%	21.7%	15.5%	13.1%	7.6%	16.7%
%*halfdayK*	43.3%	41.7%	46.3%	48.0%	40.7%	43.7%	45.3%
%*center*	71.6%	73.7%	71.0%	77.5%	78.7%	84.9%	76.9%
%*minority*	59.7%	58.3%	48.5%	33.3%	34.2%	33.9%	38.8%
wksesl							
M	−0.6133	−0.2705	−0.1234	0.149	0.2807	0.6148	0.1235
(*SD*)	(0.67)	(0.64)	(0.71)	(0.75)	(0.70)	(0.75)	(0.76)
plageent							
M	65.6	65.1	65.5	66.1	66.5	67.1	66.1
(*SD*)	(4.40)	(4.34)	(3.97)	(4.00)	(4.07)	(3.86)	(4.06)
%*public*	98.5%	93.5%	86.9%	76.7%	70.9%	61.9%	77.7%

ECLS-K 抽样方案设计要求对亚裔和太平洋岛国背景的儿童过度抽样，目前它包括三轮数据，分别在入园时、幼儿园年级结束时、一年级结束时各收集一次。三年级的数据在 2004 年春季发布。它还对 30%的进入一年级时的儿童子样本收集数据。这里所呈现的例子所使用的数据都包括在 30%的一年级子样本中；在兴趣变量上没有缺失值的儿童，也是第一次的幼儿园就读者（没有留级生），并且留在与他们所就读的幼儿园同一所学校的一年级。考虑到本书的关注点及对亚裔/太平洋岛国裔的过度抽样，以及其他少数族裔的稀少，于是对这些例证模型生成一个种族/民族的二分变量，其分类是 $1=$ 少数群体，$0=$ 白人/高加索人。具有这些特征的儿童共来自 255 所学校（57 所私立，198 所公立）共 $n=3365$ 人，平均每所学校 13 人。第 7 章将会介绍嵌套设计与定序结果数据分析的合并；所有其他的分析都假定学校之间的儿童是独立的。

第 3 章 | 背景:logistic 回归

logistic 回归概述

定序回归模型与二分结果的 logistic 模型关系紧密，所以作者先简要回顾一下 logistic 回归分析，以突出后面内容的类似性和差异性。其他 QASS 系列[①]以及其他作者(Cizek & Fitzgerald, 1999; Hosmer & Lemeshow, 1989, 2000; Menard, 1995, 2000; Pampel, 2000)已经深入讨论了 logistic 回归，所以这里仅探讨那些对本书稍后章节有重要作用的概念。

拟合定序回归模型的术语和估计策略是对那些在 logistic 回归中所使用的术语和估计策略的直接扩展。这些模型一起被定义为一般线性模型，由三个部分构成：

- 随机部分，这里因变量 Y 遵循指数族(exponential family)的某一分布，例如正态分布、二项式分布或者逆高斯分布；
- 线性部分，它描述了因变量 Y 的函数 Y' 是如何取决于一系列预测变量的；
- 关联函数，它描述了从因变量 Y 到 Y' 的转换(Fox, 1997)。

恒等关联函数(identity link function)没有改变因变量，从而得到连续结果变量的一般线性模型，原因在于多元线性回归是熟悉的案例。logit 关联函数将结果变量转换成比数的自然对数形式(下面给出解释)，这就得到 logistic 回归

① 指 SAGE 出版社的"社会科学的量化应用系列"丛书(Quantitative Application in the Social Science)。——译者注

方程。

二分结果的 logistic 分析试图对事件的发生建模，并且估计自变量对这些比数的作用。一个事件的比数是简单地将事件发生的概率（称之为“成功”）与事件未发生的概率（称之为“失败”）相比得到的商。当成功的概率大于失败的概率时，比数就比 1 大；如果两个结果可能相等，比数则为 1；如果成功的概率小于失败的概率，则比数比 1 小。

对于上面描述的 ECLS-K 数据，假设我们感兴趣的是研究进入一年级时的儿童们取得阅读熟练度 5 级（常用字）。结果可以用二分描述：一个取得了 5 级熟练度的儿童（成功）或者没有（失败）。取得 5 级的比数通过在样本数据中将取得 5 级的概率（得分为 $Y=1$）除以未能取得 5 级的概率（得分为 $Y=0$）：

$$\text{比数} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$$

为了检验自变量对比数的影响，例如性别或者年龄，我们构建了比数比，它将解释变量取不同值时的比数相比。例如，如果我们想比较男性（编为 $x=1$）和女性（编为 $x=0$）达到熟练度 5 级的比数，则用下面的比例计算：

$$\text{比数比} = \frac{\dfrac{P(Y=1 \mid x=1)}{1-P(Y=1 \mid x=1)}}{\dfrac{P(Y=1 \mid x=0)}{1-P(Y=1 \mid x=0)}}$$

比数比的下界为 0 但是没有上界；也就是它们可以从 0 到无限大。比数比为 1 时表明解释变量对成功的比数没有作用，即男性成功的比数与女性成功的比数是一样的。小的比数比值（小于 1 的）表明成功分母中的 x 值（$0=$ 女性）所代表的人群成功的比数比那些分子中（$1=$ 男性）的更大的 x 值所代表的人群成功的比数更大。大于 1 的比数比的值反过来解释也成立。也就是说，男性处于熟练度 5 级的比数比女性的同一比数要大。自变量使用的编码的本质和类型在解释中很重要；在该例以至全书中，作者使用简单的虚拟或参照编码（simply dummy or referent coding）。其他分类自变量的编码方法可以改变模型中变量作用的解释；对 logistic 回归模型中定性数据类别化的替代方法的讨论可以参照霍斯

默和莱默苏(2000)。

比数比是分类结果和一个自变量之间关联的测量,提供了"一个呈现出的结果如何随着讨论中的变量而变化的风险的明确指示"(Hosmer & Lemeshow, 1989:57)。尽管事件的概率可以直接通过线性概率模型来建模(例如,使用对二分[0, 1]因变量的普通线性回归),但这样的方法会引起一些严重的解释问题。线性概率模型可能会得到[0, 1]之外的不现实的概率预测,尤其是当自变量为连续变量时。此外,普通线性回归模型中通常的同方差性和误差正态性假设在结果变量为二分时也会违反,使得从这种方法中得到的结果的有效性受到质疑(Cizek & Fitzgerald, 1999; Ishii-Kuntz, 1994; O'Connell, 2000)。因而当结果为二分时,我们对比数建模,或者对比数的自然对数(底为 e)建模,称做 logit 分布。

这种对比数的简单转换有许多理想的特点。首先,它消除了内在于比数比估计中的偏态(Agresti, 1996),该比数范围从 0 到无穷,其值为 1 时表明比数没有变化(no change in the odds)的虚无案例(null case)。logit 的范围从负无穷到正无穷,消除了比数比和概率的界限问题。转换模型在参数上是线性的,这意味着解释变量对比数的对数的作用是附加的。因此,模型很容易处理并允许相当直接地解释变量,还允许模型建构策略与普通线性回归中所使用的那些策略类似。

该过程可以扩展到包括更多的自变量。如果我们令 $\pi(Y=1 \mid x_1, x_2, \cdots x_p) = \pi(\underline{x})$ 代表"成功"的概率,或者兴趣结果(例如,儿童处于熟练度 5 级),对于给定的有 p 个自变量的集合,logistic 模型可以写成:

$$\ln(Y') = \text{logit}[\pi(\underline{x})] = \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right)$$
$$= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

在该表达式中,Y' 仅是一种简便的方式,指转化过的结果变量的比数;而非直接预测 Y,我们预测的是 $Y=1$ 的比数(的对数)。关联函数描述了将原始的 Y 与转化后的结果 $f(y) = \ln(Y') = \ln[\pi(\underline{x})/(1-\pi(\underline{x}))]$ 相"关联"起来的过程,称为 logit 关联。解 $\pi(\underline{x})$ 我们则会得到 logistic 回归模型成功模型类似的表

达式：

$$\pi(\underline{x}) = \frac{\exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}{1+\exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}$$

$$= \frac{1}{1+\exp[-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)]}$$

SPSS 和 SAS 之类的统计软件包提供模型中变量的截距和回归参数(weights)的极大似然(ML)估计。极大似然估计通过使用得到"'最好地'解释了观测数据的总体参数值"(Johnson & Wichern, 1998:178)的迭代方法推导出来。这些极大似然估计最大化了获得原始数据的可能性，而且因为 logistic 模型是通过一个对结果的非线性转化得到的，该方法并不要求误差项的正态分布，而普通最小二乘估计要求这样。似然(likelihood)代表了观测结果可以从一系列自变量集中预测的概率。似然可以从 0 到 1 不等；似然对数(Log-Likelihood，简称 LL)也从负无穷到 0 不等。将 LL 乘以 −2 得到一个可以用来验证假设进而以比较不同模型为目的的值(Hosmer & Lemeshow, 2000)。

模型拟合的评估

评估一个拟合模型是否能够良好地再造观测数据的一种方法是，计算拟合模型的偏差(deviance)。偏差代表的是模型如何糟糕地再造了观测数据，它是通过将拟合模型与完美拟合模型——也称为饱和模型[6]——的似然相比较得到的。饱和模型的参数个数与自变量的值一样多；饱和模型的似然是 1，−2LL(饱和模型)是 0。任何模型的"偏差"，D_m 因此就是 −2LL 的量(Hosmer & Lemeshow, 2000)。我们期望更好的模型拟合的"劣度"(poorness)降低(直到 0)。两个嵌套模型的拟合，模型 1 中的变量是模型 2 自变量的子集，可以通过比较它们偏差的差异：$G = D_{m1} - D_{m2}$。G 的量代表拟合的"优度"，对于大样本，G 服从近似的卡方分布，自由度等于模型 1 和模型 2 的参数个数之差。统计上显著的 G 值意味着模型 2 拟合的"劣度"要小于模型 1。

当模型 1 是虚无模型时，这个比较提供了一个综合检验(omnibus test)(假

定是大样本属性且没有稀少的单元格），以检验拟合模型是否比虚无模型，或者是截距模型，更好地再造了观测数据。然而，它并没有告诉我们该模型与饱和模型，或者是完美模型相比表现如何。对于分类预测变量，SAS 通过皮尔森 χ^2 标准和偏差 χ^2 来检验 D_m（将拟合模型与饱和模型相比）。当模型中有连续解释变量时，这些办法都不再合适（Allison, 1999; Hosmer & Lemeshow, 2000）。当解释变量为分类时，这些测试可以在 SAS 的模型命令中通过使用"/aggregate scale=none"选项来实现。

当小样本或者数据中有稀少单元格出现（当模型中包括了连续自变量时，这种情况经常发生）时，应该考虑替代的评估模型拟合的办法；一个通常的策略是 Hosmer-Lemeshow（H-L）检验（1989, 2000）。H-L 检验在 SAS 中通过模型命令中的"/lackfit"选项来实现；在 SPSS 中，通过在打印语句中使用"goodfit"将会得到该检验。

H-L 检验在自变量（Independent Variables, IVs）是连续变量时表现很好，因为它在数据中直接处理协变量模式的数目。当自变量是连续的时，本质上数据集中的每个观测都有一个可能的不同协变量模式。简言之，H-L 检验形成若干个基于样本估计概率的称之为"风险十分位数"（deciles of risk）的群体。在大多数情况下，形成 $g = 10$ 个群体，但也许会更少，这取决于不同协变量模式间估计概率的类似性。这些分位数的案例可以用来生成一个 $g \times 2$ 的观测到预期的频次表，以及该表的皮尔森 χ^2 系数（Hosmer & Lemeshow, 1989, 2000）。如果模型拟合良好，观测和预计频次间应该达成一致，所以来自模型的在观测和预计频次间的良好拟合的虚无假设将得到保持。H-L 检验在一些文献中受到效能不够的批评（Allison, 1999; Demaris, 1992），但是依靠一个单独的检验表明模型充足性的做法本身是不应该鼓励的（Hosmer & Lemeshow, 2000）[7]。本章稍后将探讨包括关联测量和预测效率在内的更多策略。

模型的解释

通常，SPSS 对因变量中编码较高的值（如果结果编码为 0 或 1 的话则是 1）

的比数对数进行建模，但是SAS默认的是对较低的响应值进行建模。结果变量为二分时，结果和自变量对比数的作用的解释不受"成功"与"失败"如何编码的决定的影响，因为这两种事件是互补的。例如，如果令"成功"的概率或者 p(达到熟练度5级) $= 0.2$。那么"失败"的概率或者 p(未能达到熟练度5级) $= 1 - 0.2 = 0.8$。成功的比数是0.25(0.2/0.8)。互补事件的比数，即未能达到熟练度5级，将是1/0.25或者4.0(0.8/0.2)。因为对于因变量仅有两种可能结果，互补事件的比数就是该事件的倒数。当应用logistic转化时，我们可以看到取事件比数的对数与该事件的互补的比数的对数($\ln(4) = +1.3863$)有着相反的符号($\ln(0.25) = -1.3863$)，但是量级相同。在logistic回归模型中，将建模的结果的编码颠倒相当于考虑回归系数的方向和解释后与之相同的预测概率和解释。结果变量为二分时，在SAS命令中使用"降序"选项改变了默认方法并要求计算机对更高值的结果类别进行建模，即如果结果编码为0或1时，对标签 $Y = 1$ 的类别建模(或者结果编码为1时对类别2建模)。然而，定序因变量类别多于2个时，应用"降序"选项将会极大地改变模型，因此需要谨慎使用。第4章将会详细介绍对定序结果中该选项的使用。

对于 $j = 1$ 到 p 的自变量，多变量logistic模型的回归参数代表了 X_j 每增加一个单位时logit发生的变化。因为以比数而非比数对数的形式考虑变量作用更直观(回归参数是比数对数的形式)，比数自身的信息通过将模型中变量的参数指数化得到(例如 $\exp(b_j)$)。回归参数的指数化形式就是比数比，而且在计算机运行中通常都报告。比数比可以直接解释为自变量对成功的比数的作用，而且比数的百分比变化可以通过使用以下公式计算得到：($100 \times [$比数比$-1]$)。

自变量和结果变量之间的强烈关联通常可以以远大于1的比数比表现出来，不管方向如何。朗(Long，1997)将比数比称为"因素改变"(factor change)估计。对自变量一个单位的改变，相对应的比数比就是"成功"的比数预计将要改变的因素。比数比统计上的显著通常通过检验回归系数 β_j 是否在统计上显著不等于0，其实现的方法有三种：Wald检验、计分检验或者似然比检验。在Wald检验中，logistic回归各个自变量作用的参数估计除以它对应的标准误，将

其结果再平方就得到一个无作用的零假设下服从 1 个自由度的卡方分布的值。然而，在小样本中使用 Wald 统计量可能会成问题；同样，在包含多种数据模式的样本里，例如当自变量是连续的而非分类的；或者分类自变量的单元格稀少时(Jennings, 1986; Menard, 1995)，使用 Wald 统计量都会有问题。SPSS 和 SAS 针对拟合模型都会报告每个变量的 Wald 卡方统计量。

模型中一个自变量的贡献的计分检验取决于似然函数的衍生品，而且在 SAS 和 SPSS 中都不能直接得到；然而，SPSS 的确在分段过程中使用计分检验，以决定变量何时进入还是离开一个发展中的模型(Hosmer & Lemeshow, 2000)。似然比检验已经被看做是自变量对模型贡献的最可靠检验而被提倡，但在 SPSS 或者 SAS 中也没有直接提供。该测试可以很容易地通过一些简单但费时的编程实现，它涉及将嵌套模型间的偏差进行比较，也就是不含兴趣自变量的模型的偏差与包含它们的模型的偏差相比较。偏差的差别近似是 1 个自由度的卡方分布。因为本书的焦点是二分模型的发展和总体解释，作者选择了依靠 Wald 检验来评估解释变量的作用。然而，研究者还是需要意识到也存在该检验的替代方法。

关联的测量

尽管梅纳德(Menard, 2000)和其他人(Borooah, 2002; Demaris, 1992; Long, 1997)都指出，在研究者间关于哪个消减误差比例最有意义的问题上尚不能达成一致，但 logistic 回归中确有一些与来自于熟悉的普通最小二乘回归的模型 R^2 的类似物，它们对告知自变量群体与结果之间的关联强度可能有用。对于 logit 类型的模型，似然比 R^2 值，R_L^2，似乎提供了多变量模型相对虚无模型(null model)(仅有截距)拟合度改善的最直观测量。R_L^2 是通过比较两个对数似然比得到的：$R_L^2 = 1 - \text{log-likelihood}(模型)/\text{log-likelihood}(虚无)$ (Hosmer & Lemeshow, 2000; Long, 1997; McFadden, 1973; Menard, 2000)。它测量的是通过使用自变量集而得到的消减误差比例(对数似然比)(相对虚无模型)。也有其他关联强度的测量，但下面将只讨论几个。朗(1997:102)写道："当拟合

测量提供了一些信息时，它仅是部分的信息，因此必须放在激发该分析的理论背景、过去的研究以及考察的模型的参数估计的环境中进行考量。”感兴趣的读者应该参考梅纳德(Menard, 2000)关于 logistic 回归的多种 R^2 使用的讨论，以及布鲁雅(Borooah, 2002:19—23)。Huynh(2002)提供了这些情况的扩展——结果变量是定序而非二分的情况。

logistic 回归

我们将使用一个简单的例子来描述以上的概念，同时也提供一个发展中的定序回归模型的扩展。我选取以上描述的原始 ECLS-K 数据的子集：$n = 702$ 个在一年级开始时落入熟练度 0，1 或者 5 级的儿童。表 3.1 提供了这个子样本根据性别划分的频次明细。两个结果间的子样本相当平衡。在接下来的数据分析中，男性被编码为“$x = 1$”，女性被编为“$x = 0$”，结果变量中处于类别 5 级的编为“$Y = 1$”，而处于 0 或者 1 级的编为“$Y = 0$”。

表 3.1 分性别的熟练度(0，1 和 5)交互列表，$N = 702$

	$Y = 0$ (profread 类别为 0 或者 1)	$Y = 0$ (profread 类别为 5)	总计
男性 ($x = 1$)	211	151	362
女性 ($x = 0$)	134	206	340
总计	345	357	702

男性处于更高熟练度类别的比数可以通过将处于类别 5 的概率除以不处于类别 5 的概率而得到：

$$\text{比数(类别 5 | 男性)} = \frac{151/362}{211/362} = \frac{0.4171}{1-0.4171} = 0.7156$$

女性与之类似，处于熟练度类别 5 的比数可以由下式得到：

$$\text{比数(类别 5 | 女性)} = \frac{206/340}{134/340} = \frac{0.6059}{1-0.6059} = 1.537$$

从这两个值中，我们看到对于该子样本，男孩处于类别 0 或 1 而非类别 5 的概率要更大(分子小于 0.5)，对于女孩，则相反(分子大于 0.5)。因此，男孩处于类别 5 的比数比女孩处于类别 5 的比数要小。两个比数相比就得到比数比，它提供了性别和处于类别 5 的比数之间关联的测量：

$$比数比 = \frac{比数(类别 5 \mid 男性)}{比数(类别 5 \mid 女性)} = \frac{0.7156}{1.537} = 0.466$$

0.466 的比数比告诉我们，对于子样本，男孩处于更高级别的比数是女孩处于类别 5 的比数的 0.466 倍，或者说不到一半。换一种说法，作为男孩把处于类别 5 的比数降低了 53.4%(100×[比数比－1]) =－53.4)。相反的，女孩处于类别 5 的比数是男孩的比数的 2.146 倍，或者说比男孩比数的两倍还多(1/0.466 = 2.146)。

在早些时候讨论的 logistic 回归模型中，概率转化为比数，比数再通过取自然对数转换为 logit。图 3.1 给出了使用 SPSSlogistic 回归得出的以上例子的 logistic 回归模型拟合的部分结果(附录 A 中的句法，A1 部分)。在该模型中，处于类别 5 中 Y 被编码为 1，否则为 0。解释变量，“性别”编码为男孩“1”，女孩“0”。我们令 $\ln(Y')$ 代表 logit，或者比数对数。预测模型是 $\ln(Y') = 0.430 + (-0.765) \times$ 性别。参数估计在图 3.1 的最后一部分“方程中的变量”得到。

当儿童是女孩时(性别 = 0)，截距代表的是对比数的对数的预测，为 0.430。将该值指数化回比数，我们有 exp(0.430) = 1.537，它是以上所解出的女孩处于熟练度 5 级的比数。对于男孩(性别编码为 1)，我们的模型预测变成：$0.430 + (-0.765 \times 1) = -0.335$。将该结果指数化，得到 exp(−0.335) = 0.7153，这是(允许舍入误差)男孩处于熟练度 5 级的比数。最终，比数比(将舍入考虑进来)可以通过将性别的回归参数指数化得到，exp(−0.765) = 0.466。该值在“方程中的变量”表中最后一列给出，它正是由频次数据得出的比数比。它告诉我们男孩处于熟练度 5 级的比数是女孩比数的 0.466 倍。

对于诸多研究者，解释比数比比解释 logit 要容易些，但是 logit 也可以被直接解释。logistic 回归模型中性别的作用告诉我们，当性别的值变化一个单

logistic 回归

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	702	100.0
	Missing Cases	0	0.0
	Total	702	100.0
Unselected Cases		0	0.0
Total		702	100.0

注：a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0.00	0
1.00	1

Block 1: Method=Enter

Iteration History[a, b, c, d]

Iteration		−2 Log-likelihood	Coefficients	
			Constant	gender
Step 1	1	947.829	−0.755	
	2	947.825	−0.765	
	3	947.825	−0.765	

注：a. Method: Enter.

b. Constant is included in the model.

c. Initial −2 Log-Likelihood: 972.974.

d. Estimation terminated at iteration number 3 because parameter estimates changed by less than 0.001.

Omnibus Tests of Model Coefficients

		Chi-Square	df	Sig.
Step 1	Step	25.149	1	0.000
	Block	25.149	1	0.000
	Model	25.149	1	0.000

Model Summary

Step	−2 Log-likelihood	Cox & Snell R Square	Nagelkerke R Square
1	947.825	0.035	0.047

Classification Table[a]

Observed			Predicted		
			CUMSP2		Percentage Correct
			0.00	1.00	
Step 1	CUMSP2	0.00	211	134	61.2
		1.00	151	206	57.7
	Overall Percantage				59.4

注：a. The cut value is 0.500.

Variables in the Equation

		B	S.E.	Wald	*df*	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	Gender	−0.765	0.154	24.690	1	0.000	0.466	0.344	0.629
	Constant	0.430	0.111	15.014	1	0.000	1.537		

注：a. Variable(s) entered on step 1: gender.

图 3.1 SPSSLogistic 回归案例的部分输出结果

位时，预计 logit 将会发生的变化，在本例中则是从 0(女性)到 1(男性)。基于 Wald 标准，logit 模型中性别的作用统计上显著：Wald 的 $\chi_1^2 = 24.690$，$p < 0.000$。这意味着性别的估计斜率是−0.765 且在统计上显著不等于 0，比数比 $= \exp(-0.765) = 0.466$ 因而统计上显著不等于 1。

在 SPSS 例子中，虚无模型的偏差通过图 3.1 的区块 1(Block 1)“迭代历史”的脚注：$D_0 = -2LL_0 = 972.974$ 而得到。仅包括性别变量的拟合模型的偏差是 $D_m = -2LL_m = 947.825$。这两个偏差的差异是 $G_m = 25.149$，$df = 1$，

$p = 0.000$。该例的模型中仅包括一个变量时，性别作用的综合检验也是似然比检验（Wald χ^2 检验的替代）。在“模型系数的综合检验”部分得到的综合检验，意味着我们在性别包括进模型时在 $-2LL$ 中可以得到一个统计显著的减少。该缩减表明了一个可以通过似然比 R_L^2：$1-(D_m/D_0) = 0.0258$ 表示出来的削减偏差比例。在本例中，性别包括进模型后将虚无模型的偏差 $(D_0 = -2LL_0)$ 减小了 2.58%。

SPSS 和 SAS 在它们的 logistic 回归过程中都不报告 R^2，但是如上所述，它可以很容易地通过从各自统计软件已有统计量中计算获得。两种统计软件都报告 logit 分析的 R^2 统计量的两种变式：Cox 和 Snell R^2（SAS 中报告为一般化的 R^2）和 Nagelkerke R^2（SAS 中称做最大尺度的 R 平方）。Nagelkerke R^2 将 Cox 和 Snell R^2 的值重新调节以获得为 1 的界限。对于这些数据，图 3.1 中“模型概要”报告了 $R_{CS}^2 = 0.035$ 和 $R_N^2 = 0.047$。尽管综合检验在统计上显著，但没有一个 R^2 非常大，表明除性别之外的其他一些解释变量对理解儿童处于熟练度 5 级的可能性也许有用。梅纳德（Menard, 2000）探讨了若干将来自普通线性回归的熟悉的 R^2 一般化的尝试，但是他提倡将 R_L^2 作为已有虚拟 R^2（pseudo R^2's）中最有用的。

减小拟合评估中单个值的努力，如同虚拟 R^2 们所做的那样，在比较竞争（嵌套）模型时也许有些价值，但是这仅提供了一个“模型是否充分的粗略指标”（Long, 1997:102）。模型充分度的考察可以通过评估观测分类结果再生成而得以增加，基于个体是否被预测为落入他或者她原始的结果 $Y = 0$ 或 $Y = 1$ 中。该预测效率的评估补充了来自于模型拟合检验和偏差统计缩减中的已有信息。一些拟合或者在观测与预测结果之间的对应的测量，受到在结果频次分布方面高度不均衡的数据的强烈影响，所以一项信息充分的决策最好是通过计算并比较若干不同测量而非仅依靠单独一个测量来做出的。

为了考察一个模型将个案正确归类的能力，分类是基于从模型中估计的概率，而且结果是与各个类别的观测频率进行比较。对于任一儿童，如果基于 logistic 模型的“成功”的概率大于 0.5，预测结果将会是 1；否则预测结果将会是 0（Hosmer & Lemeshow, 2000; Long, 1997）。SPSS 直接生成分类表，在区块 1

中“分类表格”(Classification Table)标题下。SAS 中可以要求生成预测概率(SPSS 中也一样)进而构建分类表格;观察附录 A 中 A1 和 A2 部分的句法,介绍了如何生成这些预测概率。尽管存在许多不同的分类统计量(Allison, 1999; Gibbons, 1993; Hosmer & Lemeshow, 2000; Huynh, 2002; Liebetrau, 1983; Long, 1997; Menard, 1995, 2000),但看起来其中一些也可以用到定序因变量上。它们包括 τ_p,即“根据分类的基准率来调整预期的误差个数”(Menard, 1995:29),以及调整计数的 R^2 或者 R^2_{adjCount}(adjusted count R^2 or R^2_{adjCount}),它与古德曼—克鲁斯凯(Goodman-Kruskal)λ 系数的非对称形式类似(即当一个变量从其他一系列变量中预测时);R^2_{adjCount} 调整了个案出于偶然而被指派到观测因变量(DV)模型类别的可能概率的原始百分比校正测量(Liebetrau, 1983; Long, 1997)。不幸的是,文献中即使同样的测量也经常有若干不同的名字,读者应该密切关注各种文章或者文本中不同的术语。例如,梅纳德(Menard, 1995, 2000)就把 R^2_{adjCount} 称为 λ_p。

霍斯默和莱默苏(2000)指出,以观测和估计概率之间对应的方式的模型拟合通常比基于分类的拟合评估更可靠和有意义。它们建议分类统计可以用来作为其他测量的附属,而非一个模型质量的单独指标。以上所述,考察拟合充分性的多个标准都在这里的例子中得以描述和报告。

SAS 和 SPSS 都不直接提供 τ_p 和 λ_p(R^2_{adjCount}),但是它们可以通过分类表计算出来。为了得到 τ_p,首先需要得到 2 × 2 表格的误差期望个数,它是:

$$E(\text{error}) = 2 \times \frac{f(Y=0) \times f(Y=1)}{n}$$

要求的关联的测量进而可以通过下式得到:

$$\tau_p = \frac{E(\text{errors}) \times O(\text{errors})}{E(\text{errors})}$$

观测误差是分类表的非对角线的元素。τ_p 的另一个不同表达式可以在梅纳德(2000)中找到;它对定序因变量模型也是合适的:

$$\tau_p = 1 - \frac{(n - \sum_i f_{ii})}{\sum_i \frac{f_i(n - f_i)}{n}}$$

这里 i 代表结果变量各个类别的指数，n 为样本规模，f_{ii} = 正确预测的类别总和(在分类表的对角线上)，f_i = 类别 i 的观测频次。对于这些数据，$\tau_p = 0.1878$，表示在调整了基准率(base rate)后，将仅使用性别作为模型的唯一预测指标的分类误差缩减了大约 19%。

为了得到分类表的 $R^2_{adjCount}$ 或者 λ_p，按照朗(Long，1997)和梅纳德(2000)的方法使用下面的算式：

$$\lambda_p = 1 - \frac{n - \sum_i f_{ii}}{n - n_{mode}} = \frac{\sum_i f_{ii} - n_{mode}}{n - n_{mode}}$$

这里 n_{mode} 是指结果变量的模态类别(modal category)中观测响应的频率(最大的行边际)。对于这些数据，当观测类别被作为因变量处理时，$\lambda_p = 0.1739$。对于以上例子中建构的模型，一旦考虑因变量的边际分布的话，基于性别的预测熟练度类别属性(0，1 vs 5)则将误差减少了 17.4%。

SAS 在 LOGISTIC 过程中产生了若干关联的定序测量，可以用来补充虚拟 χ^2 和由分类表格得到的预测效率统计量，例如 Somer's D，一个等级定序统计量(Cliff，1996；Liebetrau，1983)。大部分的等级定序统计量是基于调和(concordant)与非调和(discordant)配对(pairs)的概念。"配对"是指将各个个案(个体)与每个数据集中的其他个案(不包括它本身)进行配对。对于样本量为 n 的样本，有 $n(n-1)/2$ 个可能的个体间配对。兴趣是配对的个体并不在相同的观测响应里；我们忽略那些两个案例的兴趣结果都为 0 或者都为 1 的个案。如果两个个案有不同的响应，观测值是 1 的个案的配对观测概率(基于模型而被归类到"成功")大于观测值是 0 的个案的配对观测概率，则称为调和配对；否则称为非调和配对。一个既不能归入调和亦不能归入非调和的配对(不同的响应)将被混淆(这在预测概率非常接近时可能发生；SAS 将预测概率按照 0.002 的间距长度进行类别化)(SAS，1997)。变量作用就是计算精确预测了不同结果的

各个个体配对的方向的次数。Somer's D 也许是最广泛使用的现有的等级序列相关统计量：Somer's D = (nc—nd)/t；这里 nc = 调和配对的个数；nd = 非调和配对的个数。使用 SAS 时，该例的 Somer'D 是 0.189，表示观测结果和预测概率间对应关系的强度。[8]

比较统计软件间的结果

为了方便不同统计软件间 logistic 分析的应用和解释，同时也为了将我们的讨论导向定序结果的处理，我们对先前的模型也通过 SAS 的 LOGISTIC(升序和降序两种方法)和 SPSS PLUM(定序结果)来拟合。结果的概要在表 3.2 中给出(这些模型的句法在附录 A 中，从 A1—A4)。所有的模型都使用 logit 关联函数。

表 3.2 将二分结果的 SPSS、SAS 和 SPSS PLUM 的结果进行比较：分性别的阅读熟练程度(0, 1 vs 5)[a], $N = 702$

	SPSS Logistic and SAS(descending)	SAS(ascending)	SPSS PLUM
Probability estimated	$P(y=1)$	$P(y=0)$	$P(Y\leqslant 0)$
Intercept	0.430	−0.430	0.335
gender=1(male)	−0.765**	0.765**	0
gender=0(female)			0.765**
Model fit			
−2LL(intercept only)	972.94	972.974	972.974[b]
−2LL(model)	947.825	947.825	947.825
$\chi_1^2(p)$	25.149(<0.0001)	25.149(<0.0001)	25.149(<0.0001)
Model predictions($\hat{p}$)			
Male	0.417	0.583	0.583
Female	0.606	0.394	0.394

注：a. 如果响应熟练度是 0 或者 1 的话，$Y = 0$；如果响应度是 5 的话，则 $Y = 1$。
b. 在 SPSS PLUM 的打印命令中使用"kernel"，可以求得似然的饱和值(full value)。
** $p < 0.01$。

观察表 3.2 中第 1 列的结果，注意到 SPSS 的 LOGISTIC REGRESSION 和

SAS 的 LOGISTIC 过程(降序)都是基于估计 $P(Y=1)$并拟合相同的模型,也即儿童的响应是熟练度 5 级的概率。这两个相同模型的概率预测可以通过先使用提供的估计来计算男孩和女孩的 logit,再将这些 logits 指数化以决定各个群体的比数,然后将这些比数转化成界定为“成功”的响应的概率 $p=$[比数(成功)/(1+比数(成功)]。

第二个模型的结果在表 3.2 的第三列呈现,使用 SAS 的升序选项,简单地对儿童熟练度响应类别为 0 或者 1 的概率进行建模,而非对儿童响应类别为 5 的概率进行建模。注意,截距和性别作用的符号与那些第 2 列中的相反,但仍然是同样的量级。第 2 列和第 3 列中男孩的概率估计的总和等于 1,女孩的也类似。降序选项的 SAS(第 2 列)是对默认方法(升序,第 3 列)中的事件的互补(complementary of the event)进行建模。因此,用升序方法推导出来的概率是那些在使用 SAS 降序选项得到的概率的互补(complementary probabilities)。

使用 SPSS PLUM 的模型参数估计看起来与先前使用的方法很不一样,但是实际上概率估计与第 3 列中的那些(而且正因为如此,互补的法则可以用来得到第 2 列中的概率估计)是一样的。SPSS PLUM 是一款专门设计用来分析定序因变量变量的程序,参数估计的结果将不会与那些在 SPSS logistic 回归中得到的完全对应,尤其是在 SPSS PLUM 中估计的概率处于或者低于某特定结果值的响应的概率,即更小的响应代码时;比较而言,SPSS LOGISTIC 是对较高的结果值类别的概率进行建模。此外,SAS 将二分的和定序的响应都通过它的 LOGISTIC 过程来处理,SPSS PLUM 过程使用了一种与一般线性模型稍微不同的公式:$\ln(Y_j')=\theta_j-\beta_1X_1$。在该式中,下标 j 表示响应类别,X_1 是指单个因变量,性别。性别作用的估计是从截距中减去的。另一个 PLUM 的结果与那些在 SPSS 或者 SAS 中的 logistic 回归程序的重要区别是,PLUM 在内部设置分类预测变量的编码。在第四列中,性别作用的估计对应于性别 $=0$,也即女性。所用的编码系统清晰地在输出结果中呈现(PLUM 和 SAS 的定序模型的输出结果的例子将在下一章呈现)。为了得到女孩处于(最多)熟练度 0 或 1 级的估计概率 $P(Y\leqslant 0)$,在本例中也等于 $P(Y=0)$,因为没有小于 0 的响应,我们使用该估计来得到女孩的预测 logit 值 $(0.335-0.765=-0.43)$,指数化这个结

果以得到女孩处于(或者低于) $Y = 0$ 的比数 ($\exp(-0.43) = 0.65$)，然后再解出估计概率 ($0.65/(1+0.65) = 0.394$)。使用相同的过程以得到男孩处于(或者低于)类别 0 或 1 的估计概率，或者 $P(Y = 0)$。

除了 R^2_{cs} 和 R^2_N 之外，SPSS PLUM 还提供了 R^2_L，称为麦克法登的虚拟 R^2 (Long, 1997; Menard, 2000)。为了得到 −2LL 偏差统计量的必需值，必须在 SPSS PLUM 的"/print"命令中写明"kernel"选项，可以在附录 A 中的 A4 部分看到。

关于 SAS 和 SPSS 如何处理二分结果的以上讨论及简单比较说明，尽管模型参数估计在表面上也许不同，但从模型估计中得出的结果预测概率，以及模型预测统计量，在软件和方法间是一致的。这些简单的例子也说明，分析者需要意识到，一旦选择某种统计软件，则预测的结果和分类自变量是如何整合进模型中的这一点很重要。方法和软件间的区别随着超过二分案例的定序因变量类别的增加而变得更加重要。

第4章 | 定序结果的累积(比例)比数模型

累积比数模型概述

对于仅有两个类别的结果变量，logistic 回归被用来对其中之一的结果进行建模，通常称之为"成功"，作为一系列自变量的函数。兴趣响应的估计概率，P(成功)及其互补，$1-P$(失败)，可以通过第三章中所示的例子用 logit 预测模型来得到。当一个结果变量的可能响应由多于两个的类别组成并且本质上是定序的时候，"成功"的概念可以由许多不同的方式构成。定序因变量的回归模型正是为这种情形设计，而且是二分数据的 logistic 回归模型的扩展。拟合定序回归模型的复杂性部分来自于如何才算"成功"以及对"成功"的后续概率构造模型有多种不同的可能性。

例如，给定 K 水平的定序因变量变量，例如在 ECLS-K 研究(表 2.1)中的早期阅读熟练度 $K=6$ 的，我们可以得到若干不同的"成功"代表，这取决于如何看待数据。总体而言，K 个水平的定序数据可以被 $K-1$ 个"成功"切割点分割(Fox，1997；McCullagh & Nelder，1983)。当然，成功是相对项；通常，它指定了我们关注的事件。例如，"成功"也许被定义为有儿童在熟练度上得分为"0"类，即那些不能识别大写或者小写字母的儿童。在数据的这种分割下，我们的兴趣在于找出那些与处于该最低类别的可能性增加相关联的因素，而非那些高于 0 的类别，从 1 级到 5 级。也许有某些儿童的、家庭的或者学校的有害特点与处于最低类别的可能性增加相关联。对于这些解释变量，我们计算的是处于(或者低于)类别 0 的比数。

接下来我们将"成功"设想为由处于或者低于类别 1 的事件组成；我们对这

种数据划分的兴趣在于找出相对于超过最低阶段的可能性(处于类别 2 到 5)而处于类别 0 或 1 的更大可能性相关联的因素。我们可以继续以这种累积方式描述数据,即将“成功”最终概念化为处于或者低于类别 K,这当然会一直发生。因此,数据的最后分割或者区分不再必要。使用该累积过程,在有 $K=6$ 个定序因变量类别的情形下,我们可以有关于数据的 $K-1$ 或者 5 个不同的可能“成功”事件。

模拟这种二分化结果的方法,以相继的二分化形式形成数据的累积“分割”的分析,被称为比例或者累积比数(CO)(Agresti, 1996; Armstrong & Sloan, 1989; Long, 1997; McCullagh, 1980; McCullagh & Nelder, 1983)。这的确是将数据相继地区分成二分群组的一种方法,同时仍然利用了响应变量的顺序。该方法的定序本质是如此具有吸引力,因为它具有 Logistic 回归的简洁性。如果一个单独的模型可以用来估计在所有的累积分割间处于或者低于某给定类别的比数的话,则该模型将会提供比对应于以上描述的顺序分割的数据的 $K-1$ 个不同的 Logistic 回归模型拟合大得多的简洁性。累积比例模型的目标是同时考虑一系列自变量在这些可能的相继累积分割的数据中的作用。然而,也有其他的方法定义“成功”。每种用于定序回归的不同方法都以截然不同的方式将数据区分,因此它们提出的研究问题也是不同的。数据如何分割以对应累积比例模型的概念,以及本书中将要讨论的用于拟合定序回归模型的其他两种方法——连续比例(CR)模型和相邻类别(AC)模型,都在表 4.1 的提示列中呈现。后两种方法将在后面的章节里详细讨论。本章集中讨论 CO 模型。

当应用定序回归模型时,可以从数据中得出一个简化的假设,即比例的或者平行的比数假设。该假设是指在表 4.1 中呈现的各个模型的类别(CO, CR, AC)的解释变量对比数有相同的作用,而不管数据的不同相继分割是怎样的。例如,如果用一系列对应于以上描述的 CO 模型的单个的二分 logistic 回归来拟合数据的话,平行性的假设意味着我们会观察到一个变量在所有的回归中具有共同的比数比(或者作用);即假设一个自变量对比数的作用在对应的分割间是不变的(Agresti, 1989; Brant, 1990; Menard, 1995; Peterson & Harrell, 1990)。因此,一个模型就将足够描述定序因变量和一系列预测值之间的关系。

SAS 和 SPSS 都在他们的定序回归过程中提供比例比数假设的计分检验，但是这个比例综合测试不是一个功效强大的测试且不够保守(anticonservative)(Peterson & Harrell, 1990)。测试几乎都得到非常小的 p 值，尤其是当解释变量的个数很多(Brant, 1990)，样本量很大(Allison, 1999; Clogg & Shihadeh, 1994)或者在连续的解释变量被纳入模型(Allison, 1999)时。因此，仅基于计分检验而得出拒绝比数比例性假设的结论要很谨慎，拒绝考察中的特定定序模型的平行性(比例比数)假设，意味着至少有一个解释变量也许在结果的各水平上有不同的作用，也就是说，存在着在一个或者更多个自变量与从数据中推导出的分割之间的交互项(Armstrong & Sloan, 1989; Peterson & Harrell, 1990)。问题的关键是能够找出哪个(或哪些)变量对于拒绝总体检验起作用。

一个考察自变量作用在累积 logit 中是否相对稳定的合理策略，是比较所考察的对应于定序模型(表 4.1 所示)的各个单独 logistic 回归模型间的变量作用。尽管简化比例性假设在对数据拟合总体模型时也许有用，但推荐的做法是，研究者应当检验隐含的二分模型以辅助关于定序方法适合性的决定(Brant, 1990; Clogg & Shihadeh, 1994; Long, 1997; O'Connell, 2000)。在对应的单

表 4.1 与三种不同定序模型方法相关联的类别比较，基于一个六水平的定序结果 ($j = 0, 1, 2, 3, 4, 5$)

累积比数(升序) $P(Y \leqslant j)$	累积比数(降序) $P(Y \geqslant j)$	连续比例 $P(Y > j \mid Y \geqslant j)$	相邻类别 $P(Y = j+1 \mid Y = j$ 或 $Y = j+1)$
类别 0 相对所有之上的类别	类别 5 相对所有之下的类别	类别 1 到 5 相对类别 0	类别 1 相对类别 0
类别 0 和 1 的组合相对所有之上的类别	类别 5，4 相对所有之下的类别	类别 2 到 5 相对类别 1	类别 2 相对类别 1
类别 0，1 和 2 的组合相对所有之上的类别	类别 5，4 和 3 相对所有之下的类别	类别 3 到 5 相对类别 2	类别 3 相对类别 2
类别 0，1，2 和 3 的组合相对所有之上的类别	类别 5，4，3 和 2 相对所有之下的类别	类别 4 和 5 相对类别 3	类别 4 相对类别 3
类别 0，1，2，3 和 4 的组合相对类别 5	类别 5，4，3，2 和 1 相对类别 0	类别 5 相对类别 4	类别 5 相对类别 4

独 logistic 模型拟合中，非正式的斜率比较提供了关于数据平行性可信度方面的支持信息。本章稍后将介绍一种对一些解释变量放宽比例比数假设的方法，如偏比例比数(Partial Proportional Odds，PPO)模型(Ananth & Kleinbaum，1997；Koch，Amara & Singer，1995；Peterson & Harrell，1990)。

单个解释变量的累积比数模型

为了描述累积比数模型的使用，作者从拟合仅有一个类别解释变量——性别的简单模型开始。表 4.2 给出了男孩和女孩分别在五个早期阅读熟练度类别中的频次。数据在熟练度类别的分布上不平衡，大部分的儿童，无论性别，都落在熟练度 3 级内。数据的该特性在决定最好代表了数据的模型(CO，CR，AC 或者其他)时是一个重要的考虑因素；然而，为了教学的目的，我们在这里先忽略该数据的特性，然后在不同的定序模型描述完后再检验它的影响。

累积比数模型用来预测处于或者低于某特定类别的比数。因为有 K 个可能的定序结果，模型实际上做出 $K-1$ 个预测，每个对应于相继类别间的累积概率。如果我们令 $\pi(Y \leqslant j \mid x_1, x_2, \cdots x_p) = \pi_j(\underline{\mathrm{x}})$，代表某响应落入小于或者等于 j 类的类别里 ($j = 1, 2, \cdots K-1$)，于是我们得到每一个个案的累积概率集合。最终类别的累积概率一直是 1。(注意，在 ECLS-K 数据里，作者使用 0 指代第一类，$K =$ 第 6 类指熟练度 5 级)通过一般 logistic 回归模型的扩展，累积概率的 logit 预测，被称为累积 logits，如下所示：

$$\ln(Y_j') = \ln\left(\frac{\pi_i(\underline{\mathrm{x}})}{1-\pi_i(\underline{\mathrm{x}})}\right) = \alpha_j + (\beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p)$$

可以将与处于或者低于某特定类别 j 相关联的累积 logits 指数化，以得到累积比数的估计并用来计算与处于或者低于类别 j 相关联的累积概率。

表 4.2 也以男孩女孩处于或者低于类别 j 的实际概率(p)，累积概率(cp)和累积比数(CO)的形式反映了 ECLS-K 数据的交互表。粗体行包括了这些数据的关联比数比(男孩对女孩)。最后两行提供了不分性别的累积比例 $P(Y_i <$ 类

表 4.2 性别与无核熟练度类别的交互分类观察数据：频次(f)，比例(p)，累积比例(cp)，累积比数[a](co)以及比数比(OR)

类别	0	1	2	3	4	5	总计(f)
男性							
f	48	163	320	735	256	151	1673
p	0.0278	0.0974	0.1913	0.4393	0.1530	0.0903	1.000
cp	0.0278	0.1261	0.3174	0.7567	0.9097	1.000	—
co	0.0295	0.1443	0.4650	3.110	10.074	—	—
女性							
f	19	115	274	747	331	206	1692
p	0.0112	0.068	0.1619	0.4415	0.1956	0.1217	1.000
cp	0.0112	0.0792	0.2411	0.6826	0.8782	0.9999	—
co	0.0113	0.086	0.3177	2.1506	7.21	—	—
OR	**2.6106**	**1.6779**	**1.4636**	**1.446**	**1.3972**	—	—
总计(f)	67	278	594	1482	587	357	3365
cp_{total}	0.0199	0.1025	0.279	0.7195	0.8939	1.000	—

注：a. 累积比数 = 比数($Y_i \leqslant$ 类别 j)。

别 j)。从表格中我们发现，随着响应值的增加，处于或者低于任意特定类别的比数也在增加。这在直观上说得过去，因为在样本中很少有儿童处于最高的类别里；儿童更有可能处于或者低于某给定类别而非超过该类别。通常，样本中男孩的比数大于女孩的比数，因为从比例上来讲，样本中的男孩比女孩一年级测试开始达到更高类别的比例更小。比数比让这种模式更清晰。男孩处于或者低于某特定类别的比数大约在 1.72(平均值)。在一年级开始时的熟练度定序测量中，趋势是女孩更有可能超过男孩。

与第 3 章中的例子类似，作者准备呈现使用三种不同的方法拟合一个变量的简单 CO 模型的结果：SAS 的 logistic 过程、SAS“降序”的 logistic 过程，以及 SPSS 的 PLUM(所有模型的句法在附录 B 部分提供)。表 4.1 给出了简单的一个变量累积比数模型的 SAS 输出结果(以默认的“升序”选项)。累积比数模型的合适关联函数是 logit 关联。CO 模型的另外两种方法的句法在 B2 和 B3 部分。尽管当需要 CO 模型时，从预测的角度讲这些方法在本质上是相同的，但在 CR 和 AC 定序回归模型中不一定如此。在以最简单的 CO 模型案例开始时，明

白程序和方法间的类似和差别很重要。

在使用 SAS(升序)的情况下，比数是基于较低次序类别的累积，即关联的预测累积概率对应于表 4.1 中第一列的模式。SAS 估计的是 $P(Y \leqslant 类别\ j)$，对于这些数据是 $P(Y \leqslant 0)$，$P(Y \leqslant 1)$，$P(Y \leqslant 2)$，$P(Y \leqslant 3)$，$P(Y \leqslant 4)$，当然最终的类别是 $P(Y \leqslant 5) = 1.0$（通常不包括在这些分析的输出结果上）。一个可靠的 CO 模型能够重新得到(reproduce)从表 4.2 的数据中得出的累积比数和累积概率。

在这些模型里，性别对于女孩来说编为 0 而男孩编为 1。观察图 4.1 中提供的输出结果，我们可以看到比例比假设对这些数据是成立的(比例比数假设的计分检验)，$\chi_4^2 = 5.3956$，$p = 0.2491$。我们可以总结说：性别的作用在数据的五个分割间统计上差异不显著；这表明如果用五个单独的二分 logistic 模型来拟合表 4.1 中对应的模式，则各个模型中性别的斜率(以及比数比)都将类似。因此，性别的比数比可以仅通过一个模型来同时估计。因为性别是这里唯一包含的模型，结果也告诉我们表 4.2 中的五个比数比在统计上差别不显著，一个共同的比数比可以用来概括性别对熟练度的影响。

虚拟 R^2 统计量在输出结果的“模型拟合统计量”(model fit statistics)中可以得到(图 4.1)，在“logistic 过程”(the LOGISTIC procedure)线以下，Cox and Snell $R_{cs}^2 = 0.0110$，并且 Nagelkerke(SAS 中称做最大尺度的 R 平方) $R_N^2 = 0.0116$。似然比 $R_L^2 = 0.0037$ 可以通过使用仅含截距的模型的 $-2LL$ 统计量加上概要表里“模型拟合统计量”中的模型协变量信息得到。从总体上看，这些 R^2 统计量表明响应和预测变量之间的关系很微弱。然而，模型整体拟合的检验(“检验全局虚无假设”)在统计上全部显著，而它是用来评估拟合模型相对虚无(截距模型)模型是否改善了预测，所以我们拒绝虚无模型而支持含有性别作为预测变量的模型。尽管虚拟 R^2 很低，似然比检验表明从模型中预测得到的男孩和女孩的累积比例模式(见表 4.3，稍后解释它们)比我们所期望的不考虑性别的(表 4.2 最后一行)累积比例提供了比实际中男孩女孩的累积比例(见表 4.2)更好的配对。简单 CO 模型清楚表明了这些比例在男孩女孩间是如何不同的。

The LOGISTIC Procedure

Model Information

Data Set	WORK. GONOMISS
Response Variable	PROFREAD
Number of Response Levels	6
Number of Observations	3365
Model	cumulative logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	PROFREAD	Total Frequency
1	0.00	67
2	1.00	278
3	2.00	594
4	3.00	1482
5	4.00	587
6	5.00	357

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion(GCONV=lE−8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
5.3956	4	0.2491

Model Fit Statistics

Criterion	Intercept OnlY	Intercept and Covariates
AIC	10063.980	10028.591
SC	10094.586	10065.319
−2 Log L	10053.980	10016.591

The LOGISTIC Procedure

R-Square	0.0110	Max-rescaled R-Square	0.0116

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.3884	1	<0.0001
Score	37.2553	1	<0.0001
Wald	37.2060	1	<0.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-square	Pr > ChiSq
Intercept 0.00	1	−4.1049	0.1284	1022.263	<0.0001
Intercept 1.00	1	−2.3739	0.0667	1266.52	<0.0001
Intercept 2.00	1	−1.1474	0.051	505.4293	<0.0001
Intercept 3.00	1	0.759	0.0485	245.3247	<0.0001
Intercept 4.00	1	1.9545	0.0627	971.9783	<0.0001
GENDER	1	0.3859	0.0633	37.206	<0.0001

Odds Ratio Estimates

Point 95% Wald

Effect	Estimate	Confidence	Limits
GENDER	1.471	1.299	1.665

Association of Predicted Probabilities
and Observed Responses

Percent Concordant	29	Somers's D	0.079
Percent Discordant	21.1	Gamma	0.159
Percent Tied	49.9	Tau-a	0.058
Pairs	4110137	c	0.540

图 4.1 SAS 累积比数模型例子:性别

表 4.3 预测累积 logit,男孩女孩处于或者低于类别 j 的估计比数,估计累积概率(cp),以及从 CO 模型中得到的估计比数比(SAS 的升序选项)

比　较	($Y\leqslant 0$)	($Y\leqslant 1$)	($Y\leqslant 2$)	($Y\leqslant 3$)	($Y\leqslant 4$)
男孩					
累积 logit	−3.719	−1.988	−0.7615	1.1449	2.3404
累积比数	0.02427	0.13696	0.467	3.1421	10.385
$\hat{cp}_b$	0.0237	0.1205	0.3183	0.7586	0.9122
女孩					
累积 logit	−4.1049	−2.3739	−1.1474	0.759	1.9545
累积比数	0.0165	0.0931	0.3175	2.1363	7.0604
$\hat{cp}_g$	0.0162	0.0852	0.241	0.6811	0.876
OR	**1.4711**	**1.4711**	**1.4709**	**1.4708**	**1.4709**

输出结果的下一部分(图 4.1)包括“极大似然估计的分析”,一个含五个截距的表格,被称为阈值参数(threshold parameters):$K-1$ 个切割点各一个。把这些阈值看成儿童有可能会被预测到更高类别的标记点(以 logit 的形式)很有用,但是它们通常不是单独解释的,类似于在定序多元回归模型中的截距函数。然而,通过性别的虚拟编码(性别 $=0$ 表示女孩),这些阈值估计代表了对应于 $Y\leqslant$ 类别 j 的女孩的预测 logit。性别对 logit 的作用是 0.3859,关联的比数比是 1.471[$\exp(0.3859)=1.471$]。模型告诉我们,男孩处于或者低于类别 j 的比数大约是女孩比数的 1.471 倍,不考虑我们考察的是哪种累积分割。该结果可以与我们在表 4.2 中观察数据的模式进行比价,表 4.2 中的类别间的平均比数是 1.72。根据模型,男孩相对女孩更不可能超过某一特定类别,这与实际数据一致。该模型假设性别的作用在不同的单独累积分割中是一致的。因为我们在性别作为预测变量纳入模型中时没有拒绝比例比数假设,CO 模型表明累积分割(表 4.2)的单独比数比与从 CO 模型中得到的 1.471 的比数比在统计上没有差别。

再来看模型参数估计的直接解释,截距和性别的作用可以用来估计累积比数,即男孩和女孩处于或者低于某一特定类别的比数。它们还可以用来估计各个分割的比数比,尽管我们已经从我们的分析中知道它是 1.471。对男孩和女孩的累积比数估计可以与那些从原始数据中得到的(表 4.2)进行对

比。性别 = 0,即女孩对应于各个累积类别截距的预测,被指数化后就提供了女孩处于或者低于类别 j 的响应的比数。男孩的预测通过将性别 = 1 的值代入累积比数模型的各个单独方程等式并指数化后得到比数：$\ln(Y'_j) = \alpha_j + 0.3859$(性别)。例如,对于代表 $Y \leqslant 0$ 的 logit,对女孩的 logit 预测是 −4.1049;对男孩,预测概率是 −3.719。表 4.3 提供了基于模型的这些估计累积 logits,还有男孩和女孩的估计累积比数(co)[exp(累积 logit)]。通过这些累积比数,男孩和女孩的比数比可以通过各个类别轻易得到,这些在表 4.3 的最后一行呈现(例如,$co_{男孩}/co_{女孩}$)。将误差舍入,这些比数比都近似等于 1.47。估计累积比数通过 $cp = co/(1+co)$ 转换成估计累积概率(cp),得到 $P(Y <$ 类别 $j)$。结果在表 4.3 中展现,可以与表 4.2 中观察到的累积概率进行比较。总体来看,估计值似乎很好地匹配了数据;该模型的似然比检验统计上显著。

比数比固定,并因此在所有累积类别间保持一致,这意味着总体上男孩处于或者低于任何类别 j 的比数是女孩处于或者低于类别 j 的比数的 1.47 倍。对于本例,男孩比女孩更有可能处于或者低于任一给定类别;女孩更有可能比男孩处于更高的类别。性别(男 = 1)对累积概率有着正面的作用($b = 0.3859$),对应于男孩相对女孩处于或者低于类别 j 的更大比数。这个最后解释与以这种方法进行建模的转换过的结果变量一致(处于或者低于类别 j 的响应),因而对 logit 的方向和解释变量作用的说明就依赖于这些结果是如何塑造的。

估计和实际累积概率的差别归因于 CO 模型将一个非常特殊的结构置于数据之上的事实。这个结构通过比数比的行为得以证实,因此它也影响了从模型中估计得到的累积比例。累积概率的估计是在比例比数假设下推导出来的。尽管我们稍早看到该假设在性别间成立,但认识到模型估计和预测概率是由该假设推导出的还是很重要。当那些该假设不成立或者看起来在经验上或者理论上不可信时,这些估计概率可能会明显不精确。不幸的是,当模型变得更复杂时,对比例比数假设将很难充满信心。作者将再次回到这个话题以结束本章。

该分析的 Somers'D 是 0.079(见图 4.1 的最后一部分),这是相当低的。由于仅有一个预测变量,我们在儿童配对中的预测概率的定序方向上得到的线索相当微弱。为了构建预测效率测量的分类表,τ_p 和 λ_p,我们可以使用各个儿童的累积预测概率的集合来评估个体类别归属的概率在何处达到最大。通过在输出子命令中使用"predprobs = cummulative"选项,SAS 生成一个包含了第 i 个儿童在各个层次上的累积概率的文件,$cp0_i = P$ (处于或者低于 0 级),$cp1_i = P$ (处于或者低于 1 级)等等。因此,数据集中每个儿童都有 K 个新观测,所有儿童的 $cp5_i = P$ (处于或者低于 5 级) = 1。类别概率可以通过使用关系式 $P(Y = 类别\ j) = P(Y \leqslant 类别\ j) - P[Y \leqslant 类别(j-1)]$,即 $P_i(Y = 0) = cp0_i$;$P_i(Y = 1) = cp1_i - cp0_i$;$P_i(Y = 2) = cp2_i - cp1_i$ 得到,等等。某个儿童的最大类别概率对应于他或者她的最好的熟练度预测水平。在仅有性别的升序模型里,所有的儿童都被预测到 3 级,考虑到模型的拟合很差且 3 级的儿童在整个样本中占了 44% 的事实就不会惊讶了。按照第 3 章勾画出来的方法,分类估计可以与观测熟练度类别一起做成表格以测量预测效率。对于该分析,τ_p 和 λ_p 分别是 0.23 和 0。这些结果强调了当评估模型合理时,需要连同似然比检验来考虑关联的若干不同测量。

表 4.4 提供了刚刚描述的基于以定序因变量的顺序为默认选项的"升序"SAS 模型的结果及其"降序"结果,与 SPSS PLUM 结果和仅有性别作为预测变量的多元回归模型结果之间的比较。尽管 CO 模型本质上都是相同的,提供的也是性别作用相同的解释,但还是有必要指出这些模型的呈现结果中的一些重要异同。

首先,与 SAS 对二分结果使用升序和降序选项的 logistic 过程的结果类似,阈值(截距)参数的估计在符号上相反,量级上(magnitude)却并非如此;他们在输出上也显出相反的顺序。这也许仅是因为升序和降序选项预测的是互补事件的缘故。通过降序选项,模型估计的是(相反的)累积比数,即 $P(Y \geqslant 5)$,$P(Y \geqslant 4)$,$P(Y \geqslant 3)$,$P(Y \geqslant 2)$,$P(Y \geqslant 1)$,当然,$P(Y \geqslant 0)$ 将一直等于 1。

其次,比例比数假设的计分检验表明,比例性假设在分析中成立,$\chi^2_4 = 5.3956$,$p > 0.05$,尽管 SPSS 将这称为"平行线检验"。9 对于所有的三个模

表 4.4 使用 SAS(升序),SAS(降序),SPSS PLUM,以及多元线性回归对定序因变量变量的累积比数模型:分性别的熟练度($j = 0, 1, 2, 3, 4, 5$),$N = 3365$

	SAS(升序)	SAS(降序)	SPSS PLUM	SPSS REGRESSION
模型估计	$P(Y \leqslant 类别\ j)$	$P(Y \geqslant 类别\ j)$	$P(Y \leqslant 类别\ j)$	$E(Y \mid X)$
截距 α				3.108
阈值	$\alpha_0 = -4.105$	$\alpha_5 = -1.955$	$\theta_0 = -3.719$	
	$\alpha_1 = -2.374$	$\alpha_4 = -0.759$	$\theta_1 = -1.988$	
	$\alpha_2 = -1.147$	$\alpha_3 = 1.147$	$\theta_2 = -0.762$	
	$\alpha_3 = -1.955$	$\alpha_2 = -2.374$	$\theta_3 = 1.145$	
	$\alpha_4 = 1.955$	$\alpha_1 = -4.105$	$\theta_4 = 2.340$	
性别 = 1(男性)	0.386**	−0.386**	0	−0.246**
性别 = 0(女性)				
R^2	0.004[a]	0.004[a]	0.004[a]	0.012
计分检验[b]	$\chi_4^2 = 5.3956$	$\chi_4^2 = 5.3956$	$\chi_4^2 = 5.590$	
	($p = 0.2491$)	($p = 0.2491$)	($p = 0.232$)	
Model fit[c]	$\chi_1^2 = 37.388$	$\chi_1^2 = 37.388$	$\chi_1^2 = 37.388$	$F_{1,3363} = 40.151$
	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)

注:a. R_L^2 = 似然比 R^2。
b. 针对比例比数假设。
c. 定序模型的似然比检验;针对普通最小二乘法(OLS)回归的 F 检验。
** $p < 0.01$。

型,综合似然比检验表明定序性别模型比虚无模型拟合得更好,$\chi_1^2 = 37.388$, $p < 0.001$。

第三,使用 SAS 升序和 SPSS PLUM 的累积比数和累积比例的预测完全一样;并且 SAS 降序得到的累积比数预测得到的是这些概率的互补。回忆下 SPSS PLUM,模型预测通过从阈值估计中减去性别作用的影响可以得到。SPSS PLUM 也使用一套内部的编码系统来预测类别。例如,为了使用 PLUM 模型估计一个女孩的熟练度响应低于或者等于 2 级的累积概率,我们将(a)得到 $\ln(比数比(Y \leqslant 2)) = \theta_2 - \beta_{(性别=0)} = -0.762 - (0.386) = -1.148$;(b)指数化以得到比数,$\exp(-1.148) = 0.3173$;(c)使用这些比数得到女孩的累积概率,$P(Y \leqslant 2) = 0.3173/(1 + 0.3173) = 0.2409$,与表 4.3 中 SAS 升序结果一

致。为了阐明 SAS 降序选项的方法，考虑 $Y \leqslant 2_{女孩}$ 的互补，即 $Y \geqslant 3_{女孩}$。使用表 4.4 中降序模型的参数估计，我们得到累积对数 $\log_{女孩,Y\geqslant 3} = \alpha_3 + (-0.386) \times$ 性别 $=+1.147$（因为女孩的性别 $=0$）。进而得到累积比数 $odds_{女孩,Y\geqslant 3} = \exp(1.147) = 3.149$。估计概率是 $P(Y \geqslant 3)$，女孩 $= 3.149/(1+3.149) = 0.759$。这是使用 SAS 或者 SPSS PLUM 得到的表 4.3 中 $P(Y \leqslant 2)$ 的互补概率：$1-0.2410 = 0.759$。

第四，如前所述，所有的程序都被要求保存估计概率，它们可以很容易地（至少对那些预测变量数据很少的模型）与那些定序数据进行比较。SAS 在运行 CO 模型时，将根据操作者的要求（升序还是降序）计算并保存累积概率。然而，SPSS PLUM 并不能直接得到累积概率，相反，计算并保存的是个体类别属性的概率。如上所述，累积概率可以直接由三个模型的任意一个参数估计信息得到。

三个模型的解释是相同的，尽管阈值和截距的实际值在 SPSS 和 SAS（升序或者降序）之间并不一致。这可以简单地归因于两种软件参数化所拟合模型的方式不同。SAS 升序和降序方法的差异已经在符号相反的标记阈值的下标中看到。降序方法的累积比数是处于或者超过某特定熟练度水平的比数；升序方法和 PLUM 的累积比数是处于或者低于某特定熟练度水平的比数。两种 SAS 方法输出结果中阈值看起来是相反的顺序，但是一旦预测 logit 被转化成了累积概率，其结果在本质上是相同的。两种 SAS 模型中性别的作用符号相反，在 PLUM 中性别作用对应于性别 $=0$ 的个案，但是有兴趣的读者在将模型特征和推导的累积概率也考虑在内的情况下，有必要用这些简单模型来验证预测概率的均等性。接下来是三种模型间性别处理的一个例子。

在 SPSS PLUM 中，阈值估计是针对性别 $=1$（即男性）时的个案，而在 SAS 中，阈值估计针对的是性别 $=0$（女性）的个案。不考虑所用的分析，性别的作用在所有的累积分割间的作用是一致的，$b=\pm 3.86$。例如，使用 SAS（升序），男孩处于熟练度 2 级或者更低的 logit 预测是 $\alpha_2 + b_{性别(男孩们)} = -1.147 + 0.386 = -0.761$。这与那些根据 SPSS PLUM 分析得到的男孩的预测一样：$\theta_2 - b_{性别(男孩们)} = -0.762 - 0 = -0.762$。指数化以得到累积比数，再将结果转换就得

到男孩处于或者低于熟练度 2 级的预测概率，则有 $P(Y \leqslant 2) = 0.318$（表 4.3）。在所有累积分割间基于比例比数模型的男孩对女孩的比数比假设是一致的，而且在 SAS 和 SPSS PLUM 间是等价的：$\exp(0.386) = 1.47$。这表明男孩处于或者低于任一类别 j 的比数是女孩处于或者低于类别 j 的比数的 1.47 倍。

使用 SAS(降序)方法，我们可以说男孩相对女孩处于或者超过类别 j 的比数在所有累积分割间是一致的：$\exp(-0.386) = -0.680$，这表明男孩相比女孩更不大可能处于或者超过某一给定熟练度水平。但其解释稍有不同，这个结果表明男孩处于或者超过某类别 j 的比数是女孩该比数的 0.68 倍。女孩更有可能处于较高的熟练度类别。注意，两种方法（升序和降序）的比数比是互为倒数的：$1/0.68 = 1.47$。男孩处于或者低于类别 j 的预测概率，例如，也能够从 SAS（降序）模型中得到，因为 $P(Y \leqslant j) = 1 - P(Y \geqslant j+1)$。在该过程的另一个例子中，为了得到男孩的 $P(Y \leqslant 2)$，我们可以使用降序模型来得到男孩的 $Y \geqslant 3$ 的累积 logit，$\alpha_3 + b_{性别(男孩们)} = 1.147 + (-0.386) = 0.761$；指数化并求得累积概率，得到 $P(Y \geqslant 3) = 0.682$；最终，$1 - 0.682 = P(Y \leqslant 2) = 0.318$，这与表 4.3 中所示的结果一致。

我们看到，当这些模型的结果与多元回归分析的结果进行比较时，男孩在熟练度上低于女孩的类似模式。多元回归分析的因变量熟练度被编码为从 0 到 5 的值。性别变量的斜率（男孩 $= 1$）是 -0.246。平均来看，女孩们被预测到处于 3.109 的熟练度水平，男孩们则处于低的熟练度水平 $(3.109 - 0.246) = 2.863$。整体而言，尽管定序模型和多元回归模型在性别作用的方向上有些类似，但多元回归模型的预测结果与我们分析的数据并不一致。当我们的响应值是严格的定序数据时，一个平均的熟练度水平并不是我们想要预测的；多元回归模型不允许我们做出分类的结论，以在不同的熟练度水平间进行比较。

累积比数的全模型分析

以上分析很好地表明了含一个变量的模型可以继续改善。在比例比数假设下对仅有性别的模型的预测概率与实际累积比例很相似，而且似然比测试结

果表明，当性别纳入模型时，累积概率比虚无模型(不含性别的)更多地与实际数据一致。R^2 统计量非常小，Somers'D 和预测效率测量也是。我们再来看一个更复杂的累积比数模型的推导，以得到六个熟练度水平上附加的解释变量和累积概率之间的关系。表 4.5 提供了用表 2.2 中八个解释变量来拟合 CO 模型的结果概要(公立是学校层次的变量在这些个体层次的模型中不会使用)。表 4.5 的结果是通过降序的 SAS 得到的；建模的概率是：$P(Y \geqslant 类别\ j)$。这种方法为稍后进行 CR 和 AC 定序模型间的比较提供了便利。全 CO 模型的句法包括在附录 B 的 B4 部分。

表 4.5 累积比数(CO)的全模型分析，SAS(降序)($Y \geqslant$ 类别 j)，$N = 3365$

Variable	$b[se(b)]$	*OR*
α_5	−6.01(0.54)	
α_4	−4.73(0.53)	
α_3	−2.62(0.53)	
α_2	−1.30(0.53)	
α_1	0.50(0.54)	
gender	−0.50(0.06)**	0.607
famrisk	−0.26(0.08)**	0.771
center	0.09(0.08)	1.089
noreadbo	−0.32(0.09)**	0.729
minority	−0.15(0.07)*	0.862
halfdayK	−0.17(0.07)*	0.847
wksesl	0.71(0.05)**	2.042
plageent	0.06(0.01)**	1.063
R_L^2	0.05	
Cox & Snell R^2	0.14	
Nagelkerke R^2	0.15	
Somer's D	0.33	
τ_p	0.21	
λ_p	0.00	
Model fit[a]	$\chi_8^2 = 524.17(P < 0.0001)$	
计分检验	$\chi_{32}^2 = 75.47(P < 0.0001)$	

注：a. 似然比检验。
b. 对比例比数假设。
$*p < 0.05$；$**p < 0.01$。

该模型的比例比数假设不成立，可以在表 4.5 的标有“计分检验”一行看到。这表明一个或者更多自变量的作用模式很有可能在单个的二分模型间不一样，这些模型是根据稍早在表 4.1 中 CO 模型所表明的模式拟合的。不幸的是，在连续预测变量和大样本中，计分检验几乎是一直表明应该拒绝比例比数假设的，因此需谨慎地解释(Allison，1999；Greenland，1994；Peterson & Harrell，1990)。我们将回到该假设的检验，至于现在，让我们解释一下模型估计和拟合统计量对该分析意味着什么。

模型拟合的卡方值表明，这个全模型在预测熟练度累积概率时比虚无模型(没有自变量的)表现更好。我们在似然比和虚拟 R^2 统计量上看到一些改善，但改善的没有我们通过仅有性别的模型获得的那么多。Somer's D 是 0.333，这与通过仅有性别的模型获得的值相比明显要更好。

回忆一下通过六个类别以 0，1，2，3，4，5 的结果测量的熟练度。使用降序选项，表 4.5 中的阈值估计对应于那些在自变量整个系列上得分为 0 的学生的累积 logits 预测。α_5 对应的是 $Y \geqslant 5$ 的累积 logit，α_4 对应的是 $Y \geqslant 4$ 的累积 logit 等等。直到 α_1 对应的是 $Y \geqslant 1$ 的累积 logit，因为所有的学生都有 $Y \geqslant 0$，它作为第一个阈值就没有包括在降序的 logit 模型中(请注意，对 $Y \leqslant 5$ 的升序累积 logit 模型也是如此)。

全 CO 模型里的自变量作用强调了变量是如何对处于或者超过某一特定类别的概率有所贡献的。与早些时候的仅有性别的模型一致，男孩更不大可能比女孩超过某一特定类别(比数比 = 0.607)。任何家庭风险因素的出现(famrisk，比数比 = 0.771)，父母没有对他们的孩子阅读(noreadbo，比数比 = 0.729)，少数族裔(minority，比数比 = 0.862)，以及上半天幼儿园而非全天幼儿园(halfdayK，比数比 = 0.847)在模型中都是负系数而且对应的比数比都显著小于 1。这些特征与一个儿童处于更低的熟练度类别而非更高的熟练度类别相关联。另一方面，如幼儿园的年龄(plageent，比数比 = 1.063)，以及家庭社会经济地位(wksel，比数比 = 2.042)都与高的熟练度类别相关联。这两个变量的斜率都是正的且在多变量模型中都在统计上显著不为 0。入园前参加过基于中心的日间看护在该模型中与熟练度没有关联；斜率很小，而且

比数比接近 1。这些发现与文献中关于早期阅读能力的影响因素一致,就这点而言,全模型提供了关于这些选择变量影响该领域的熟练度的方式的合理看法。

在预测效率方面,τ_p 和 λ_p 相对仅有性别的模型都没能提供更好的类别预测,而仅有性别的模型是将所有的儿童都预测到 3 级中。对于全模型 CO 分析,累积概率可以用来得到如性别模型中描述的个体类别概率,以对应于各个儿童所得到的最好的熟练度预测类别的最大可能性的类别。表 4.6 提供了基于全 CO 模型的分类方案结果。大部分的儿童仍然被归入到熟练度 3 级中,而且我们可以从分类表中(用第三章中呈现的公式)得到 $\tau_p = 0.23$ 和 $\lambda_p = 0$,表明相对于仅有性别的分析没有在预测上有总体改进。如果预测是模型的唯一目的,的确让人沮丧。然而,正如在二分 logistic 回归例子中提到的,这些测量未能告诉我们解释变量在熟练度类别间是如何影响累积概率的。霍斯默和莱默苏(2000:157)评论说,分类对群组规模非常敏感,并且"一直倾向于归入到大的群组中,该事实并不以模型的拟合为转移"。考察模型拟合中各个自变量的贡献时,应该偏好综合似然比检验(omnibus likelihood ratio test)以及 Wald 检验的结果。尽管如此,在一些研究情形下,分类的可靠性还是模型选择标准的一个重要部分;本例描述了这些统计量是如何计算的,还有他们被群体规模所影响的程度。

表 4.6 全 CO 模型的分类表,$N = 3365$

	Predcat 0	Predcat 1	Predcat 2	Predcat 3	Predcat 4	Predcat 5	总计
profread							
0	0	1	9	57	0	0	67
1	1	2	12	262	0	1	278
2	0	3	24	565	0	2	594
3	1	3	24	1428	0	26	1482
4	0	1	1	577	0	8	587
5	0	0	0	332	0	25	357
总计	2	10	70	3221	0	62	3365

logit 中的比例比数和线性假设

在定序模型中，不能直接评估 logit 的线性，并且“只有当 logits 和协变量之间的线性关系在单独的二分 logistic 模型中成立时才是比例比数假设有意义的一个验证”(Bender & Grouven, 1998:814)。因此，这里分别考察五个二分模型的(比例比数)假设，以期为定序模型提供支持。连续变量的 logit 线性的检验通过 Box-Tidwell 方法(Hosmer & Lemeshow, 1989; Menard, 1995)和图形方法(graphical methods)(Bender & Grouven, 1998)实现。对于 Box-Tidwell 方法，生成连续解释变量的乘数项 $X\times\ln(X)$ 并加入到主效应模型中。统计上显著的交互项是线性假设对该变量不一定成立的表示。为了从图形上考察线性关系，对连续变量生成十分位数，进而再与每个二分 logit 的处于“成功”类别(处于或者超过类别 j)的儿童的比例做散点图。这里同时采取了这两种方法考察模型的两个连续变量：入园年龄(plageent)和家庭社会经济地位(wksesl)。图片展示了一个线性趋势，但是 Box-Tidwell 方法表明两个连续变量在所有的五个二分 logits 中都是非线性的。考虑到做图的模式、样本量大，以及统计检验的敏感性，两个连续变量的 logit 线性被假设为是有道理的。

对于全 CO 模型，比例比数或者平行比数假设的计分检验被拒绝。这意味着有一些自变量对于处于或者超过类别 j 的比数的作用在随着熟练度 j 改变时并不稳定。表 4.7(值已被舍入以节省篇幅)提供了五个单独的二分 Logistic 回归的结果，在这里数据被二分化并对应于表 4.1 中第二 CO 列的模式展开分析。即每个 logistic 模型都考察了处于或者超过类别 j 的概率。对于这些 logistic 模型(使用 SPSS)，编码为 1 的类别归类对应于那些处于或者超过各个相继类别的儿童，编码 0 对应于那些低于各个相继类别的儿童。

回顾表 4.7 中单独 logistic 回归的结果，相对表 4.5 中 CO 模型的结果，我们可以看到，所有的五个二分模型都很好地拟合了数据。模型的 χ^2 值统计上都不显著，每个模型相对它的虚无模型都拟合得更好；H-L 检验在统计上都不显著，从观测到预测的概率是一致的。

表 4.7 CO 分析(降序)的关联累积二分模型，$CUMSP_j$ 将 $Y \leqslant$ 类别 j 与 $Y \geqslant$ 类别 j 比较，$N = 3365$

Variable	$CUMSP_1$ b [se(b)] OR	$CUMSP_2$ b [se(b)] OR	$CUMSP_3$ b [se(b)] OR	$CUMSP_4$ b [se(b)] OR	$CUMSP_5$ b [se(b)] OR	Score Test[a] P value
Constant	3.53	−0.55	−2.15**	−4.67**	−7.34**	
	(2.11)	(0.99)	(0.68)	(0.68)	(0.99)	
gender	−1.04	−0.60	−0.47	−0.46	−0.46	0.249
	(0.28)	(0.12)	(0.08)	(0.08)	(0.12)	
	0.35*	0.55*	0.63*	0.63*	0.64*	
famrisk	−0.15	−0.25	−0.21	−0.33	−0.28	0.450
	(0.29)	(0.13)	(0.09)	(0.10)	(0.15)	
	0.87	0.78	0.81*	0.72*	0.76	
center	−0.03	−0.10	0.10	0.10	0.26	0.219
	(0.28)	(0.14)	(0.09)	(0.10)	(0.16)	
	0.97	0.91	1.10	1.10	1.30	
noreadbo	−0.65	−0.36	−0.28	−0.28	−0.5	0.095
	(0.27)	(0.14)	(0.10)	(0.12)	(0.21)	
	0.52*	0.70*	0.75*	0.76*	0.61*	
minority	−0.23	−0.42	−0.39	0.09	0.24	0.000
	(0.29)	(0.13)	(0.09)	(0.09)	(0.13)	
	0.80	0.66*	0.68*	1.09	1.27*	
halfdayK	0.07	0.00	−0.11	−0.26	−0.11	0.033
	(0.26)	(0.12)	(0.08)	(0.08)	(0.12)	
	0.93	1.00	0.89	0.77*	0.89	
wksesl	1.00	0.77	0.73	0.64	0.87	0.000
	(0.17)	(0.10)	(0.07)	(0.06)	(0.08)	
	2.73*	2.17*	2.07*	1.89*	2.39*	
plageent	0.02	0.05	0.06	0.06	0.08	0.645
	(0.03)	(0.02)	(0.01)	(0.01)	(0.02)	
	1.03	1.06*	1.06*	1.06*	1.08*	
R^2_L	0.125	0.097	0.092	0.070	0.096	
R^2_N	0.136	0.128	0.149	0.115	0.128	
Model χ^2_8	82.11**	215.49**	366.40**	280.39**	217.92**	
H-L[b] χ^2_8	7.80	10.43	13.41	0.74	9.16	

注：a. 对每个自变量的计分检验，未调整(模型中没有其他协变量)。
b. Hosmer-Lemeshow 检验，均不显著。
$*p<0.05$，$**p<0.01$。

现在让我们来看各个解释变量在五个模型间斜率和比数比的模式。性别的作用在调整了其他自变量后，的确看起来在五个单独 logistic 回归分割中有着不大类似的模式。尽管性别在这五个回归中的平均斜率是－0.604，这与多变量 CO 模型中的性别斜率(－0.500)多少有些接近，男孩对女孩的处于或者超过熟练度 1 级的比数比(0.354)相对于其他四个比较(0.552、0.625、0.631 和 0.635)或多或少有些低。然而，请注意，如果将这些二分模型的平均性别斜率的比数 [exp(－0.604) = 0.547] 与来自 CO 模型的单个性别比数比(0.607)比较，我们看不到什么差异。平均来讲，性别的作用在五个 logistic 回归间类似。这对其他所有的解释变量而言都成立，除了少数族裔的作用外。注意，少数族裔变量的作用方向在前三个分析和后两个分析之间有变化。在前三个分析中，比数小于 1，表明少数族裔儿童比非少数族裔儿童更有可能处于较低的熟练度类别。然而，处于或者超过熟练度 4 级的可能性并没有差异，因为比数比在统计上没有显著不等于 1。最后的分析将从类别 0 到类别 4 的儿童与类别 5 的儿童相比较。这里我们看到，在调整了模型中其他解释变量的出现后，少数族裔儿童比非少数族裔儿童更有可能处于类别 5 中 ($b = 0.238$，比数比 =1.268)。该结果在累积比数模型中不明显。CO 模型提供了一个变量在数据的所有累积熟练度类别的二分体或者分割间的概要估计。将比数的比例性假设置于这些比例分割之上看起来对少数族裔变量无效。对在模型中的所有其他变量，斜率和比数比的方向和平均量级很好地对应了 CO 结果。

不幸的是，比例比数假设的计分检验对样本规模和不同的协变量可能模式很敏感，这就使得它在有连续解释变量的时候通常会变得非常大。如果没有拒绝假设，研究者应该很自信地说整体 CO 模型很好地代表了在单独累积分割间的比数比模式。然而，如果假设不成立，合适的做法是拟合单独的模型并将其与 CO 结果比较以检查与由 CO 模型表明的通常模式下的区别或者偏差。例如，Allison，1999；Bender & Grouven，1998；Brant，1990；Clogg & Shihadeh，1994；Long，1997；O'Connell，2000。

为了提供比例性假设真实性的更多检验，可以观察未经根据累积比数模型中其他协变量的出现调整的对各个解释变量的单独的计分检验。考虑到样本

量大，所以使用 0.01 的显著性水平来指导关于非比例性(nonproportionality)的决定。对于每一个单独的二分模型，比例比数假设的计分检验成立，但少数族裔和家庭社会经济地位例外。表 4.7 的最后一栏给出了这些未经调整的检验的 p 值。在五个二分 logit 模型中，wksels 的比数比接近 1.9 或者更大，表明高家庭社会经济地位的儿童处于较高熟练度类别的可能性是较低社会经济地位的儿童的两倍。考虑到社会经济地位是连续变量，比数比在二分分割间的差异的量级看起来是可以忽略不计的，就其本身而言，一个共同的比数比对该变量也许是更为正当的假设。然而，如上所述，少数族裔比数比改变的模式将会清晰地与熟练度研究相关，该变量的作用应该更仔细地加以考察。尽管不在这里提供，后续包括了预测变量之间的交互项或者使用一个种族/民族的单独类别的变量而非一个总体指派的少数族裔类别的分析将能够更好地解释五个二分 logit 模型的作用。

累积比数模型的替代方法

累积比数模型的最佳应用是为数据提供一个单独简洁预测模型。然而，如果均等斜率假设的约束不现实，那研究者就有责任去尽量解释数据的行为而非强行要求数据遵循某特定模型。在考察了单个的 logistic 回归分析并检查了线性和比例性后，如果多变量定序模型的比例性假设的确让人怀疑的话，那么就会产生几种替代方法。

如果变量的作用极其重要，研究者可以决定使用单独的 logistic 回归来探求并解释不同累积模型中的多种解释变量模式(Bender & Grouven, 1998)。该决定依靠研究者分析的总体目标，而且并不是对每个情境或者研究问题都合适。如果需要一个简洁模型或者单独一套预测概率，这些单独的二分 logit 将不会提供。作为替代，研究者也许要决定完全放弃因变量的定序属性而去拟合数据的一个多分类模型。该方法可以以综合变量作用和分类的形式提供一些有意义的信息，但是它忽视了结果的定序属性，因而也忽略了数据的一个重要方面。该方法对研究者的目标也不是最优化的，但如果研究者相信解释变量的大

多数都对比例比数假设的违反有所贡献的话,也可以考虑这个办法。这些替代多分类方法的例子和讨论可以参见:Borooah(2002), Ishii-Kuntz(1994),以及 Agresti(1990, 1996)。

第三个方法将在本书的稍后章节里集中讨论,它将其他类型的定序回归分析,例如连续比例方法或者相邻类别方法,尝试并获得一个单独的拟合良好且简明的模型来辅助我们取得对手头数据的理解。第 5 章描述了 CR 或者连续比例模型的使用,第 6 章呈现了 AC 模型(Adjacent Categories model),也即相邻类别模型。

在转到讨论这些分析定序结果的更多策略之前,将呈现一种方法。在仅基于解释变量的一个子集而使比例性受到质疑的情景下,研究者也可以选择用偏比例比数(Partial Proportional Odds, PPO)模型拟合(Ananth & Kleinbaum, 1997; koch et al., 1985; Peterson & Harrell, 1990)。在本质上,PPO 模型允许自变量和不同 logit 比较之间的交互项,这也阐明了一个自变量的比数是如何在比较的结果水平间变化的。SAS 当前使用 GENMOD 过程估计 PPO 模型。分析要求数据重构以反映个体是否处于或者超过某特定响应水平(Stokes, Davis & Koch, 2000)。在重构数据集中,每个人的每个定序 logit 比较都生成一个新的二分响应以表示该个体是否处于或者超过某一特定响应水平。例如,通过一个 K 个类别的定序因变量,每个人都会在重构数据集中有 $K-1$ 行数据。推导出的新的兴趣结果变量是为了表明对于 $K-1$ 个 logits 中的每一个,个体是否处于或者超过类别 K(包括最低类别 $Y=0$,所有的儿童都是超过它的)。因为数据现在是相关的(人群中的重复观测),一般估计方程(Generalized Estimating Equations, GEE)被用来拟合非线性模型进而拟合偏比例比数模型。在兴趣结果是分类的(名义的或者定序的)时候,GEE 方法(Liang & Zeger, 1986)非常适合用于对结果变量随时间推移的重复测量。它基于大样本性质,这意味着样本规模必须足够大以产生可靠的估计。斯托克斯等人(Stokes, 2000)建议,数据的二维交互分类中的观测数量应该至少达到五个。有连续解释变量的话,这通常不是问题,所以应该谨慎地考虑样本规模。

偏比例比数

通过使用 ECLS-K 数据例子来描述，我们放弃少数族裔变量的比例比数假设并尝试以一种能够更好地反映表 4.7 中的模式的模型来重新拟合。也就是说，比例比数假设在模型中除了少数族裔变量外的所有变量间都得以保持。PPO 模型以及数据集重构的句法，包括在附录 B5 中，位置在由斯托克斯(Stokes et al.，2000)等人勾勒出的过程后面。图 4.2 展现了该分析的(编辑过的)输出结果。GENMOD 对儿童处于或者超过类别 j 的概率建模，但是因为比数比在除了少数族裔外的所有各个变量的分割中都保持一致，结果仅包括了一个截距参数。阈值通过将各个对应分割的估计相加得到，这包括在图 4.2 中“GEE 参数估计分析”的中间部分。阅读这份表格时，请注意解释变量的编码方案“0”类是分类变量的参照组。例如，性别的斜率，$b = 0.5002$，是为女孩(性别 $= 0$)而非男孩(性别 $= 1$)提供的。

截距−6.9805 是表示如果他或者她的所有协变量分数等于 1，或者等于 0(如果是连续变量的话)时，儿童将会处于或者超过熟练度类别 5($Y \geqslant 5$)的比数对数；注意，分类变量的编码遵循一种内在建构的模式，而且第五个分割的估计是 0.00。为了得到 $Y \geqslant 4$ 的比数对数，阈值将是截距加上第四个分割的作用($-6.9805 + 1.2670 = -5.7135$)。其他阈值估计也可以以类似的方法得到。

GEE 分析提供了检验各个解释变量对模型的贡献的分数统计量；这些在输出结果的最后得到。计分检验的结果表明，少数族裔对第五个 logit 比较的作用仅是边际显著，$\chi_1^2 = 4.04$，$p = 0.0445$，然而它与分割变量的交互在总体上确实很显著，$\chi_4^2 = 28.80$，$p < 0.0001$。该结果表明少数族裔依赖于分割的作用存在着可靠的差异。对于模型中的其他解释变量，所有的作用都在统计上显著，除了参照日间护理中心外，这与全累积比例模型中的发现一致。GENMOD 也提供了解释变量对模型中贡献的 z 检验(Wald 统计量的正态分布版本)；这些可以在图 4.2 中“GEE 参数估计分析”表格中找到。z 检验的结果与计分检验一

The GENMOD Procedure

Model Information

Data Set	WORK. PPOM
Distribution	Binomial
Link Function	Logit
Dependent Variable	beyond
Observations Used	16825

Class Level Information

Class	Levels	Values				
split	5	1 2 3 4 5				
GENDER	2	0 1				
FAMRISK	2	0. 00 1. 00				
CENTER	2	0. 00 1. 00				
NOREADBO	2	0. 00 1. 00				
MINORITY	2	0. 00 1. 00				
HALFDAYK	2	0. 00 1. 00				
CHILDID	3365	0212014C	0294004C	3035008C	3042008C	3042023C
		0044007C	0195025C	0243009C	0621012C	0748011C
		0832023C	3041005C	0028009C	0028014C	0052003C
		0052007C	0195020C	0196007C	0196016C	0196017C
		0196018C	0212002C	0212012C	0216006C	0220005C
		0220020C	0301002C	0301004C	……	

Response Profile

Ordered Value	beyond	Total Frequency
1	1	10045
2	0	6780

PROC GENMOD is modeling the probability that beyond="1".

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	17E3	12003. 9325	0. 7142
Scaled Deviance	17E3	12003. 9325	0. 7142
Pearson Chi-Square	17E3	16074. 6352	0. 9564
Scaled Pearson X2	17E3	16074. 6352	0. 9564
Log Likelihood		−6001. 9662	

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter			Estimate	Standard Error	95% Confidence Limits		Z	Pr>|Z|
Intercept			−6.9805	0.5753	−8.1081	−5.8528	−12.13	<0.0001
GENDER	0		0.5002	0.0663	0.3703	0.6301	7.55	<0.0001
GENDER	1		0.0000	0.0000	0.0000	0.0000	.	.
FAMRISK	0.00		0.2596	0.0778	0.1071	0.4120	3.34	0.0008
FAMRISK	1.00		0.0000	0.0000	0.0000	0.0000	.	.
CENTER	0.00		−0.0759	0.0770	−0.2268	0.0750	−0.99	0.324
CENTER	1.00		0.0000	0.0000	0.0000	0.0000	.	.
NOREADBO	0.00		0.3366	0.0913	0.1575	0.5156	3.68	0.0002
NOREADBO	1.00		0.0000	0.0000	0.0000	0.0000	.	.
MINORITY	0.00		−0.1560	0.1240	−0.3989	0.0870	−1.26	0.2083
MINORITY	1.00		0.0000	0.0000	0.0000	0.0000	.	.
HALFDAYK	0.00		0.1451	0.0666	0.0145	0.2757	2.18	0.0295
HALFDAYK	1.00		0.0000	0.0000	0.0000	0.0000	.	.
WKSESL			0.7450	0.0514	0.6442	0.8457	14.49	<0.0001
P1AGEENT			0.0588	0.0084	0.0423	0.0753	7.00	<0.0001
split	1		6.2595	0.1851	5.8968	6.6223	33.82	<0.0001
split	2		4.4067	0.1204	4.1707	4.6428	36.59	<0.0001
split	3		3.0966	0.1043	2.8923	3.301	29.7	<0.0001
split	4		1.2670	0.0846	1.1012	1.4328	14.98	<0.0001
split	5	0.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	1	0.00	0.5077	0.2779	−0.0370	1.0523	1.83	0.0677
split * MINORITY	1	1.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	2	0.00	0.6021	0.1621	0.2843	0.9198	3.71	0.0002
split * MINORITY	2	1.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	3	0.00	0.5230	0.1334	0.2615	0.7844	3.92	<0.0001
split * MINORITY	3	1.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	4	0.00	0.0282	0.1050	−0.1777	0.2340	0.27	0.7884
split * MINORITY	4	1.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	5	0.00	0.0000	0.0000	0.0000	0.0000	.	.
split * MINORITY	5	1.00	0.0000	0.0000	0.0000	0.0000	.	.

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr>ChiSq
GENDER	1	57.02	<0.0001
FAMRISK	1	11.1	0.0009
CENTER	1	0.97	0.3243
NOREADBO	1	13.3	0.0003
MINORITY	1	4.04	0.0445
HALFDAYK	1	4.75	0.0294
WKSESL	1	195.03	<0.0001
P1AGEENT	1	49.02	<0.0001
split	4	2447.16	<0.0001
split * MINORITY	4	28.8	<0.0001

图 4.2 少数族裔的偏比例比数:GEE 分析

致，除了少数族裔外。

考虑到少数族裔和分割间的交互项，少数族裔的作用通过具体考察它对第五个累积比较的贡献的计分检验来解释。对于其他的各个分割，包括在模型中的少数族裔×分割的交互的 z 检验表明在第一个累积比较 ($Y \geqslant 1$) 的少数族裔儿童和非少数族裔儿童的比数没有显著差别，$b_{\text{int.}1} = 0.5077$，$p = 0.0677$，第四个也无差别 ($Y \geqslant 4$)，$b_{\text{int.}4} = 0.0282$，$p = 0.7884$。实质上，这些发现与那些表 4.7 中单独的二分模型一致。在那里，少数族裔对个体的累积 logits 没有统计作用，无论是第一个二分模型 ($p > 0.05$) 还是第四个 ($p > 0.05$) 都是这样。

少数族裔×分割的交互项告诉我们，关于少数族裔在响应变量的阈值间的作用发生了多大的改变。伴随着比例比数假设对少数族裔的放宽，输出中的结果告诉我们，非少数族裔儿童相对于少数族裔儿童的比数对数在不同的 logistic 回归分割间预计能够改变多少。例如，在调整了模型中的其他协变量之后，非少数族裔儿童相对少数族裔儿童处于或者超过类别 5 级的熟练度分数的比数比是 $\exp(-0.1560) = 0.855$；少数族裔儿童相对非少数族裔儿童处于或者超过类别 5 级的熟练度分数是 $\exp(+0.1560) = 1.169$。该比数比可以与表 4.7 中的第五个累积 logistic 回归分割相比（比数比 $= 1.268$）。该比数比在 PPO 模型中显著不等于 1（“类型 3 的 GEE 分析的分数统计量”表格中 $p = 0.0445$），实际上是基于单独二分模型的第五个累积比较（表 4.7 中最后一个分割的 $p < 0.05$）。

为了得到少数族裔对第四个累积 logit 的作用 ($Y \geqslant 4$)，将交互项加到主效应中，即非少数族裔儿童处于或者超过类别 4 的比数是 $\exp(-0.1560 + 0.0282) = \exp(-0.1278) = 0.880$；对于少数族裔儿童，则为 $\exp(+0.1278) = 1.136$。少数族裔儿童有 1.36 倍的可能性处于或者超过类别 4，尽管该作用在统计上没有明显不等于 1($p = 0.7884$)。该作用与表 4.7 中第四个累积 logit 中的比数比一致，它在统计上也不显著（比数比 $= 1.092$，不显著）。对于第一个 logit($Y \geqslant 1$)，非少数族裔儿童的作用是 $\exp(-0.1560 + 0.5077) = \exp(0.3571) = 1.42$；对于少数族裔儿童，则是 $\exp(-0.3517) = 0.7035$。根据 PPO 模型，该作用不显著 ($p = 0.0677$)，与表 4.7 中第一个分割的少数族裔的作用一致。总之，少数族裔儿童比他们的非少数族裔同龄人更不大可能超过熟练度 2 和 3 级，

但是一旦他们获得了至少 4 级的熟练度时，他们比他们的非少数族裔同龄人更有可能取得熟练度 5 级的水平。

为了检验那些保留了比例比数假设的解释变量的作用，斜率估计可以直接解释。对于性别的影响，在调整了模型中其他协变量之后，女孩(性别 = 0)处于或者超过类别 1 的可能性是男孩的 exp(+0.5002) = 1.65 倍，，而且该比数比在所有隐含累积 logit 中保持一致。因为仅有两个水平的解释变量的事件是互补的，我们也可以轻易给出对男孩作用的解释：在调整了其他协变量之后，男孩处于或者高于某一给定熟练度类别 j 的可能性是女孩的 exp(−0.5002) = 0.606 倍。对于除了少数族裔以外的其他解释变量，一旦考虑进自变量的编码，它们的作用则与表 4.5 中的全 CO 模型呈现出等价。例如，全 CO 模型中性别的斜率是−0.500，比数比是 0.607。PPO 模型中那些比例比数假设得以保留的变量的作用与那些 CO 模型中的处于相同的量级和统计显著性。方向发生了变化是因为 SAS 的 GENMOD 过程提供了解释变量编码为 0 而非 1 的值的估计。需要注意的是，分类自变量的编码实质并没有影响 CO 模型和 PPO 模型间连续变量的结果。

简要总结一下 PPO 分析，该方法的确解决了围绕全比例比数模型的一些问题，尤其是少数族裔变量。GEE 估计与在二分 logit 模型中以量级和统计显著性两种方式同时检验的少数族裔的单独作用对应得非常好。在早期阅读能力的研究中，该结果经得起进一步的考察。而另一种做法，即生成一个将儿童以种族/民族群组分类的变量并将其纳入这些模型，而非将所有“非白人”儿童放在二分法的一个赋值中，也值得进一步检验。但这不是当前叙述的重点。服从比例比数假设约束的变量的作用与早些时候的 CO 分析结果中的变量作用一致。当前还没有对由 GENMOD 提供的 GEE 分析(Stokes, et al., 2000)的拟合优度的总体概要检验，但是在“评估拟合优度的指标”标题下的输出结果中的标准表明，偏差(通过比较拟合模型与完美，或者饱和模型而得到)比它的自由度要小(值/自由度 < 1)，这暗示着模型拟合的充分性(Allison, 1999)。当模型中有连续变量时，没有模型偏差的可靠检验。然而，这些统计量对竞争模型间的比较很有用。总之，PPO 模型看起来比 CO 模型更有信息量，尤其是考虑到少数族裔变量的解释时。

第 5 章 | 连续比例模型

连续比例模型概述

如在第 4 章所看到的，累积比数模型使用所有现存的数据来评估自变量对处于或者超过(抑或相反，处于或者低于)某特定类别的对数比数的作用。比数通过考察处于或者高于某类别的概率相对于低于该类别的概率而得到。CO 分析的一个限制性假设是，在所有的累积 logit 比较中，自变量的作用是类似的；也就是说，处于较高的类别相对于低于它的任意类别的比数在不同类别间保持恒定。然而，这些累积比数的 logit 比较理论上也许并不是在每个研究情境中都合适。如果兴趣在于判断自变量对处于更高阶段或者类别的事件的作用，就包括了所有未能进入某一类别的人的比较群组也许并不能让我们达到对数据最好的结论或者理解——以一种处于低类别的和所有更高类别的人们间的差异的形式。与其将所有未能进入任一类别的人群归组到一起，不如采取一种替代的定序方法，即将那些在任意已知类别的响应者与那些取得更高类别分数的响应者进行比较。这种方法形成了称为连续比例(Continuation Ratio，CR)模型的一类模型。CR 模型分析的焦点在于理解区分那些达到了某一响应水平但是并没有继续向前的人与那些确实达到了更高水平的人之间的区分因素。福克斯(Fox，1997：472)将该过程称为一系列“嵌套二分类”(nested dichotomies)分析。

连续比例是一种条件概率。接下来的讨论将解释这些连续比例是如何根据研究者的目的而以不同方式形成的。呈现的例子是基于采取 $\delta_j = P$(超过类别 j 的响应|至少在类别 j 的响应)形式的连续比例，或者它的互补形式 $1-\delta_j = P$(处于类别 j 的响应|至少在类别 j 的响应)。

阿姆斯特朗和斯洛(Armstrong & Sloan, 1989)、麦卡拉和尼尔德(McCullagh & Nelder, 1983)、格林兰(Greenland, 1994)以及阿格雷斯提(Agresti, 1990, 1996)都已经深入讨论并强调 CR 模型和由考克斯(Cox, 1972)提出的比例风险模型之间的关系。比例风险模型在流行病学和生存分析研究文献中是常见的模型,但是它的价值可以扩展到其他情境中。

CR 模型可以通过使用适当重构的数据集进行拟合,可以是 logit 关联函数也可以是互补双对数(clog-log)关联函数。重构的细节稍后解释,其实质是,从 $K-1$ 个小的数据集中产生出一个新的数据集,其中,一个人的数据行与他或者她的结果分数一样多。该过程类似于为了偏比例比数分析而产生联结的数据集(concatenated date set)的过程,重要的一点例外是:纳入数据集的标准基于是否达到了特定熟练度水平。得到的数据集对应于表 4.1 中连续比分析的特定比较。一旦数据集联结起来,在已知已经至少取得那个类别的熟练度时,兴趣的结果就在于儿童是否超过某特定类别。以这种方式取得的数据集是条件独立的(Armstrong & Sloan, 1989; Clogg & Shihadeh, 1994; Fox, 1997),因此,重构的数据集可以通过二分结果的统计模型来分析。

为了推导出所需的条件概率,重构是必须的,在互补双对数关联情形下则是风险。在流行病学文献中,风险(hazard)比也被作为相对风险(risk);它是两个风险的比值,而风险是明确的条件比例。另一方面,比数比是两个比数的比值,而比数是互补概率的商 $p/(1-p)$。当然,logit 模型中兴趣概率也可能是条件概率,它澄清了 logit 关联对连续比模型的有用性。我们在下面的例子中可以看到,两种关联函数共享一些相似性,但是结构上却是不同的。

第一步,使用 logit 关联,被称为"logistic 连续比例模型"(Greenland, 1994: 1668)。模型将事件的次数作为离散量处理。定序结果 Y 并非必须为次数化的变量,接下来的讨论将会说明这一点。对于 ECLS-K 一年级数据,这个"次数"可以概念化为每个人在六个结果的各个不同水平来测量。让 $K=6$ 个响应类别代表超过类别 j 的事件可能会发生的"次数"。进而每个人最多可能有 $K-1=5$ 个事件发生的机会(因为没有观测到基于离散定序结果的样本中有人超过最高的类别)。对于每个人的每一次,高于类别 j 的兴趣事件,要么发生,

要么不发生。产生两个新的变量:次数——也称为“水平”或者“阶段”;以及结果——接下来的讨论中称为“超过”:要么儿童超过了特定水平(1),要么儿童没有超过特定水平(0)。通过对数据集的简单重构(下面展示)和 logit 关联的使用,这种方法的结果在概念上类似于并接近于离散比例风险模型。使用 logit 关联的 CR 模型提供了儿童超过某特定类比的比数,以已经达到或者超过那个类别为条件。与累积比数模型类似,logistic CR 模型假设可以用来呈现数据的各个嵌套二分模型的斜率是同质性的(见表 4.1)。在这种情况下,约束被称为等斜率或者平行比数假设。

第二种方法是将事件的次数作为连续量。在这种情况下,对重构数据集使用的是互补双对数关联,模型提供的是“连续次数下隐含的比例风险模型估计”(Allison, 1995:212)。模型假设风险而非比数,在结果比例的各个层次上是平行的。与比例比数假设类似,平行斜率(或者风险)假设意味着解释变量的作用被设为在结果类别间是相等的。

比例风险模型在响应变量为定序时尤其有用,因为从群组连续模型中估计的变量作用,之前考察的累积比数模型是一个例子,与使用互补双对数关联时的连续比例模型中的作用估计是等价的(Läärä & Matthews, 1985; McCullagh & Nelder, 1989)。本德和本纳(Bender & Benner, 2000:680)将群组连续模型类别定义为:

> “群组连续模型”(grouped continuous model)的名称可以以如下观点解释:Y 是一个隐含的由分割点 j 所定义的潜在连续特征的离散变量,进而再以累积概率 g_j 的形式形成模型。然而,假设一个潜在连续变量的存在进而使用累积概率来描述定序类别在本质上也并不必要。

关联函数

第 3 章描述了关联函数,采用的是将观测响应的变式(transformation of the observed respinses)与原始数据“关联”起来的形式。对于 logit 关联,作者首先将结果看成是某一类别的概率的形式,称之为“成功”。接下来,我们得到成

功的比数，最终的变换涉及取这些比数的对数形式。令 $\pi(\underline{x})$表示给定一系列协变量$\underline{X}$的情况下成功的概率，logit 关联函数可以写成：$g(y) = \ln[\pi(\underline{x})/(1-\pi(\underline{x}))]$。基于此 logit 关联，我们拟合一个线性模型，从而成功响应的 logit 预测为：$g(y) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip}$。如果使用 logit 关联，则 logistic 分布是该过程的倒数，所以：$\pi(\underline{x}) = \exp(g(y))/[1+\exp(g(y))]$，因而 logistic 回归模型提供了对成功或者失败原始概率的估计。相比其他关联函数，logit 关联通常更受青睐，因为它以比数和比数比的形式解释结果更具简洁性。

logistic 分布函数提供了自变量和二分结果之间关系的合理评估，但它不是唯一可供使用的分布。对于二分数据，其他同等效力的关联函数及其对应的分布包括 probit 关联函数以及它的倒数—累积标准正态分布，互补双对数函数及其倒数，极值分布。福克斯(Fox, 1997)提供了这些变换的大致描述，有兴趣的读者也可以参考布鲁雅(Borooah, 2002)、麦卡拉和尼尔德(McCullagh & Nelder, 1989)以获得关于这些以及其他关联方法的具体信息。

互补双对数关联函数对转换的响应以如下方式建模：$g(y) = \log(-\log(1-\pi(\underline{x})))$，这里 $\pi(\underline{x})$代表给定某系列协变量时“成功”的概率。在连续比模型中，$\pi(\underline{x})$是条件概率，即 $\pi(\underline{x})$代表的是某人一旦取得某一特定阶段后继而超过该阶段的概率，或者是它的互补，即某人在取得某一特定阶段后并没有超过该阶段的概率。基于该变换的线性模型是：$g(y) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip}$。如果使用互补双对数关联，该过程的倒数是极值分布，于是 $\pi(\underline{x}) = 1 - \exp(-\exp(g(y)))$。

相关的概率

CR 模型的兴趣概率是在已知某人已至少取得任意特定类别时，超过该类别的概率，$\delta_j = P(Y_i > 类别\ j \mid Y_i \geqslant 类别\ j)$。这些比例是条件概率，而非累积概率，预测条件概率的过程也不同于 CO 模型中使用的过程。尤其是，随着 j 的增加，所有小于 j 的响应个案都被各个 logit 比较的模型舍弃。注意，δ_j 的互补 $1-\delta_j$，等于 $1-P(Y_i > 类别\ j \mid Y_i \geqslant 类别\ j) = P(Y_i = 类别\ j \mid Y_i \geqslant 类别\ j)$。这实际上是通过比例风险模型寻求的概率。如果次数是连续的，而我们的兴趣

在于对事件发生的次数(用 T 表示)建模,则风险比例描述的是已知 T 尚未发生时,T 将在任意间距(可能会非常小)内发生的概率。如阿利森(Allison, 1995)所描述的,通过对概率求极限得到风险:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T \leqslant T \geqslant t)}{\Delta t}$$

在某种意义上,该极限在 $P(T = t \mid T \geqslant t)$ 处收敛,这也是为何风险有时候也被称为失败率。坦巴克尼克和芬戴尔(Tabachnick & Fidell, 2001:779)将风险率或者失败率描述为"未能存活到某间距中点的比例,已知在间距开始处存活"。在定序结果的情形下,存活意味着当达到 j 水平后继续朝着超过该水平的响应移动的情况,而未能生存意味着某人在达到 j 水平后在水平 j 处停止并不再向前,与生存时间不同,我们是生存事件——已知已达到类别 j 时,个人要么超过该类别 j,要么没能超过。那些未能达到类别 j 的人在早些时候就失败了,因此不会包括在超过他们最后取得的阶段的概率计算中。这些估计概率可以通过使用连续比例模型直接得到。

为了得到连续比例模型,数据必须重构以反映每个个体是否在每一个响应水平上存活。之后,发展出使用 logit 关联的 CR 模型与比例风险模型的方法是一致的。通过 logit 关联模型获得的概率估计了"正存活着"的条件概率,即已知已至少取得某特定掌握的程度,再在早期阅读连续体上取得更高阶段的概率。当对同样的重构数据集使用互补双对数关联函数时,分析等价于离散时间比例风险模型(Allison, 1999)。对于 ECLS-K 数据,通过重构数据集的互补双对数关联模型得到的概率也是条件概率,估计的是已知至少取得某给定水平的掌握度,再沿着早期阅读能力连续体水平超过更高阶段的概率。接下来的例子将会描述,使用互补双对数关联的模型斜率估计等价于在累积比数分析中对原始数据使用互补双对数关联(Läärä & Matthews, 1985)。

因变量的向性和连续比例的形成

在思考一些例子之前应该意识到,定序结果的编码方向选择很重要,因为

该选择在连续比例方法中变得极为重要。编码本身完全由研究者决定，在这个意义上只要保持方向的一致性则其他部分完全可以是随意的。例如，数字 0，1，2，…5 是用来表明是否掌握早期阅读能力(表 2.1)。也可以使用相反的情况，例如，0 表示取得 5 级的掌握度，而 5 代表未能取得 1 级的技能。早些时候我们在累积比数模型中看到，对结果的解释在相反的编码方案中仍然完全相同，尽管斜率所代表的作用的方向也是相反的(例如，与全 CO logit 模型中性别的斜率是－0.500 相反，如果结果变量使用相反的编码体系的话，则斜率将会是＋0.500)。SAS 中升序和降序选项仅改变兴趣结果建模的方式。例如，在 CO 模型中如果使用降序选项的话，我们得到 $P(Y>$ 类别 $j)$ 而非 $P(Y\leqslant$ 类别 $j)$。在 CO 模型中，变量的作用并不会因为结果如何编码而发生变化。然而，当我们开始考虑条件概率估计时，我们推导出的模型以及它们的解释则明确依赖于这些定序结果的编码方式——要么增加要么减少。只要谨慎的分析者意识到他或她想要从数据中探询的问题并对应于这些问题正确地设置定序结果的编码，那么就不会混淆。在以下例子中，这种对应将会更明显。

有若干种建构连续比例模型的独特方式，它们被称为“向前的连续比例”或者“向后的连续比例”(Bender & Benner，2000；Clogg & Shihadeh，1994；Hosmer & Lemeshow，2000)。这些特点不会通过简单地将结果类别的顺序颠倒而得到，因为结果编码颠倒后 CR 模型并非不变(Allison，1999；Greenland，1994)。这里呈现的模型使用的是向前的方法，它是与六个层级结构的早期阅读能力类别的进步过程自然对应的。尤其是当兴趣在于对已知第 i 个儿童已至少取得某一特定类别，他或她会超过该类别的概率进行建模时。对于解释变量的集合，$\underline{X}$，兴趣概率是 $P(Y_i>$ 类别 $j\mid Y_i\geqslant$ 类别 j，$\underline{x})$。注意，该事件的互补是 $1-P(Y_i>$ 类别 $j\mid Y_i\geqslant$ 类别 j，$\underline{x})=P(Y_i=$ 类别 $j\mid Y_i\geqslant$ 类别 j，$\underline{x})$。稍后将会展示这种向前连续变量的特殊构造是如何与使用原始数据的互补双对数关联分析结果相一致的。

使用 logit 关联和重构数据的连续比例模型

对于 ECLS-K 一年级的例子，我们首先以这些条件概率的形式重新考虑原始数据，如表 5.1 中提供的单个解释变量——性别。令要求的概率 $P(Y_i >$ 类别 $j \mid Y_i \geqslant$ 类别 $j,\ \underline{x})$ 用 δ_j 来表示，$j = 0,\ 1,\ \cdots 5$。已知处于某类别或者更高，再超过该特定类别的比数，通过计算 $\delta_j/(1-\delta_j)$ 得到。从表 5.1 中我们可以看到，男孩和女孩超过某一特定熟练度水平的条件概率 δ_j，通常在类别间都是降低的，尽管男孩的 δ_j 在类别 3 和类别 4 间稍微上升。从总体上看，随着取得更高的早期阅读技能，儿童愈加不大可能超过更高的类别。对于男孩，已知超过的可能性小于女孩超过的可能性。表 5.1 中男孩和女孩的比数对应于这些观测的条件概率模式。通过这些数据，女孩超过的比数，以她已经取得某给定的熟练度水平的事件为条件，总是比男孩对应的比数要大。这些数据的比数比(男孩对女孩

表 5.1　观测的 ECLS-K 性别频次(f)，类别概率(p)，以及条件概率 P(已知至少类别 j 时超过类别 j)(δ_j)

	类别						
	0	1	2	3	4	5	总计(f)
男孩	48	163	320	735	256	151	1673
f	0.0278	0.0974	0.1913	0.4393	0.1530	0.0903	1.0000
p	0.9713	0.8997	0.7811	0.3564	0.3710	—	—
δ_j	0.0287	0.1003	0.2189	0.6436	0.6290	—	—
$1-\delta_j$	48	163	320	735	256	151	1673
女孩							
f	19	115	274	747	331	206	1692
p	0.0112	0.0680	0.1619	0.4415	0.1956	0.1217	1.0000
δ_j	0.9888	0.9313	0.8241	0.4182	0.3836	—	—
$1-\delta_j$	0.0112	0.0687	0.1759	0.5818	0.6164	—	—
Odds							
男孩	33.84	8.970	3.568	0.5538	0.5898		
女孩	88.286	13.556	4.685	0.7188	0.6223		
OR	**0.3833**	**0.6617**	**0.7616**	**0.7705**	**0.9477**	—	—

的)看起来并不类似。男孩的超过比数是女孩的 0.3833 到 0.7705 倍不等,直到最终阶段(类别 4 对类别 5),两个性别超过的可能相同(比数比 = 0.9477)。

如早些时候提到的,对这些数据使用 CR 模型拟合的过程涉及将这些数据重构以使得每个儿童有他或者她的结果分数所允许的足够多的行(Allison, 1999; Bender & Benner, 2000; Clogg & Shihadeh, 1994)。从根本上说,重构的数据集包含了五个联结但不同的数据集,以某儿童是否超过了各个类别的形式代表六个熟练程度的结果。第一个数据集包括了所有的观测,那些熟练度分数超过 0 的儿童获得一个值为 1 的新变量"超过"(beyond);否则就得到一个值为 0 的"超过"变量。第二个数据集舍弃了那些没有取得至少 1 级的儿童,接下来重复以上的过程,即熟练度分数超过 1 的儿童获得一个值为 1 的"超过"变量,否则为 0。在下一个数据集中,熟练度为 1 级的儿童被舍弃(伴随着那些更早被舍弃的儿童),超过熟练度 2 级的儿童在"超过"上得分为 1,否则为 0。最终的阶段仅包括那些至少达到了 4 级熟练度的儿童,如果他们继续往前进而掌握了 5 级,他们将会在"超过"上得分为 1,否则为 0。本例中,在类别 3 上停止并出局(停止被测试)的儿童将会对联结数据贡献 4 行的信息,其"超过"分数分别是 1, 1, 1 和 0,分别代表着从 0 类到 3 类的进步过程。附录中的句法 C1 展示了如何在 SAS 中生成这些重构数据集;同等的过程也可以在 SPSS 中实现。总体上,重构数据集的总体样本规模将会对应于每一步所包括的人数。对于 ECLS-K 例子,这就是 $n = 1 \times f(0) + 2 \times f(1) + 3 \times f(2) + 4 \times f(3) + 5 \times (f(4) + f(5)) = 13053$。

"超过"变量现在是新的兴趣结果变量。我们感兴趣的是控制了阶段(或者数据集或者条件 logit 比较)时的 P(超过=1),可以通过降序选项和 logit 关联在 SAS 的 LOGISTICS 过程实现(句法 C2)。在重构句法中,"crcp"项(连续比数分割点的首字母缩略词)表明所指的是哪个 logit 比较或者数据集或者阶段(类别)。这些变量再被虚拟编码(dumcr0 到 dumcr3),在分析中将最终比较作为参照。仅有性别作为预测变量的二分 logistic 模型的概要结果在表 5.2 中呈现,该表也包括了五个仅有性别的二分 logistic 回归结果,这些回归对应于稍早在表 4.1 中呈现的各个嵌套的连续比例分割。因为这些模型本质上不同于第 4

章呈现的 CO 模型,CO 模型和 CR 的 logit 关联分析的斜率不应该拿来做比较。CR 和 CO 的 logit 关联模型预测的是非常不同的概率集。

在解释 CR 分析的结果时,首先要提出的问题是模型是否拟合。这有两部分的含义:(1)模型总体拟合,可以通过比较拟合模型和虚无模型,或者是截距模型的似然值;(2)在不同响应水平比较间考察 logit 模型的平行比数假设,或者均等斜率假设。

如表 5.2 中所见,拟合模型比虚无模型能更好地再造数据,χ_5^2 的 $=4070.84$,$p<0.0001$。这对各个单独的隐含条件 logit 模型也都成立,除了最后的比较,即处于 4 级和 5 级的儿童间的比较。因此,考察全局 logistic CR 模型的简洁性相对于用分开的模型同时拟合是否更好地代表了数据就显得很有理由了。该考察也将提供对平行比数假设的综合检验。

表 5.2 使用重构数据集的 CR 模型(logit 关联),$N=1053$;以及五个条件二分 logistic 回归模型的各个 logistic 回归结果(P(超过类别 j|响应至少在类别 j))

	CR	0 vs. above	1 vs. above	2 vs. above	3 vs. above	4 vs. above
Intercept	−0.3763**	4.478**	2.606**	1.545**	−0.330**	−0.474**
dumcr0	4.4248**					
dumcr1	2.9113**					
dumcr2	1.9283**					
dumcr3	0.0578					
gender	−0.2865**					
(OR)	**(0.751)**	**(0.384)**	**(0.662)**	**(0.762)**	**(0.770)**	**(0.948)**
Model fit						
−2LL(model)	10021.534	643.899	1896.43	2985.67	3233.103	1251.899
$\chi^2(df)$	4070.84[b](5)	13.568(1)	10.683(1)	8.830(1)	9.740(1)	0.157(1)
	$p<0.0001$	$p<0.001$	$p=0.001$	$p=0.003$	$p=0.002$	$p=0.692$

注:a. 似然比卡方。

b. 该检验是基于 $n=13053$ 个观测。

** $p<0.01$。

与偏比例比数模型类似,也可以产生自变量性别和代表连续比数分割点的各个虚拟编码变量之间的交互项。如果均等斜率假设成立,就不应期望交互项

能够必然改善模型的拟合。如果有自变量和分割点间的交互项，就表明自变量的作用在分析的分割点间不是同质性的。通常，解释变量和分割点指标编码之间的交互项能够用来产生无约束的CR模型，在其中所有解释变量的作用被允许在不同分割点间变动；抑或是产生偏CR模型，在其中平行性的假设被放宽到仅对解释变量的子集。这种方法反映了第4章中呈现的无比例或者偏比例比数模型。

再将四个交互项加到六个参数的附加模型上。交互模型的最终 $-2LL_{int}$ 是 10011.002（分析未予展示），均等模型是 $-2LL_{no\text{-}int} = 10021.534$。它们的差异是 10.532 及四个自由度（加到模型上的交互项个数）。该差异超过了临界 $\chi^2_{4,\,0.05} = 9.49$，但没有超过 $\chi^2_{4,\,0.01} = 13.28$。考虑到大样本规模，我们有理由假设交互项的一组斜率在统计上并没有显著不等于0。然而，单个二分 logistic 回归的斜率估计和比数比提供了关于性别在不同响应水平上的有用信息，这些信息通过全局或者更简洁的CR（无交互的）方法并不明显。也许我们需要额外的解释变量，但亦可能是平行比数的假设不适合数据。尽管平行比数假设在 $\alpha = 0.05$ 水平上边际显著，但为了教学的目的，我们将继续探讨与均等斜率假设下的CR模型相关联的统计量和结果。

注意，表5.2中单独的二分 logistic 回归模型的变异（$-2LL$（模型））总和与以上描述的交互的CR分析的模型变异相等（$\sum[-2LL$（模型）$] = 10011.002$）。嵌套的二分体(nested dichotomies)是独立的(Fox, 1997)，因此，"各个 G^2（变异）统计量的总和是与同时拟合模型相关的总体拟合优度统计量"(Agresti, 1996:219)。

为了更进一步考察平行比数的假设，我们需要回顾一下来自单独二分 logistic 回归的斜率和比数比，并将其与CR的结果做非正式比较。作为替代，采用一种类似于将多个模型变异相加的过程，对CR分割系列模型中的每个变量都执行将 Wald 检验统计量相加的方法(Fox, 1997)。无论采用何种方法，回顾单独分割的 logistic 回归结果总是明智的；这也是此处采用的方法。

回到表5.2的估计，logistic CR模型的性别作用是 -0.2865（$p < 0.01$），对应的比数比是 $\exp(-0.2865) = 0.751$。该比数比概述了早些时候在原始

数据中看到的趋势(表 5.1):男孩相对女孩不大可能超过某一特定的熟练度水平。为了计算条件比较的估计比数比,可以用 CR 模型来计算 logits。在所应用的虚拟编码方案中将最后的比较作为参照,我们可以用截距来得到超过熟练度 4 级的 logit,也就是:对于 $(Y>4 \mid Y\geqslant 4)$ 的 $\text{logit}_{cr4} = -0.3763 + (-0.2865 \times$ 性别)。对于男孩,$\text{logit}_{cr4} = -0.6628$;对于女孩,$\text{logit}_{cr4} = -0.3763$。通过这些值,可以得到估计的条件概率。对于男孩,$\hat{\delta}_4 = [\exp(-0.6628)/(1+\exp(-0.6628)] = 0.3401$;女孩的 $\hat{\delta}_4$ 是 0.4070。最后可以得到男孩和女孩的比数比,并计算出估计的 $OR_4 = 0.7509$。所有比较的预测在表 5.3 中概述。使用 CR 模型,所有连续比例分割点的比数比都相等并大约等于 0.75。CR 模型将这些比数在相继的条件类别比较间约束为相等。

将表 5.3 中从 CR 模型中估计的比数比与实际的性别比数比相比较,我们可以看到,定序熟练度尺度的中间具有比两端更大的相似性。尽管共同比数比是在正的方向,并且总体模型可以解释为表明男孩对于女孩更不大可能超过任何熟练度类别,这种同质性的缺乏需要更多的关注,以在模型简洁性和变量作用的明确阐述之间作出平衡的明智决定。

表 5.3　$P(Y>j \mid Y\geqslant j)$ 的观测比例(δ_j),预测和观测的以及估计的性别模型的比数比,logit 关联的 CR 分析

	类别					
	0	1	2	3	4	5
男孩						
δ_j	0.9713	0.8997	0.7811	0.3564	0.3710	
logits	3.762	2.2485	1.2655	−0.6050	−0.6628	—
$\hat{\delta}_j$	0.9773	0.9045	0.7798	0.3532	0.3401	—
$(1-\hat{\delta}_j)$	(0.0227)	(0.0955)	(0.2202)	(0.6468)	(0.6599)	
女孩						
δ_j	0.9888	0.9313	0.8241	0.4182	0.3836	
logits	4.0485	2.535	1.552	−0.3185	−0.3763	—
$\hat{\delta}_j$	0.9829	0.9266	0.8252	0.4211	0.4070	—
$(1-\hat{\delta}_j)$	(0.0171)	(0.0734)	(0.1748)	(0.5789)	(0.5930)	
OR(obs.)	0.3833	0.6617	0.7616	0.7705	0.9477	—
OR(est.)	0.7490	0.7502	0.7501	0.7508	0.7509	

从表 5.2 中条件数据集中得到的二分 logistic 回归模型可以看到，性别在各个分割间斜率估计同质性的更多证据，在量级和统计显著性两方面都是如此。对于二分模型，性别对对数比数的作用范围在 -0.054 到 -0.956 之间，平均的斜率是 -0.391。该平均斜率对应的比数比是 0.676。对于最后一个二分 logit 比较 $(Y>4 \mid Y\geqslant 4)$，性别的作用在统计上并不显著，它的比数比接近 1。尽管 CR 模型给出了性别在所有单个 logit 比较间作用的简洁描述，而且平均来讲它与单个的拟合看起来也相对应，但在使用拟合的 CR 模型时还是有一些顾虑。在一些研究情形下，呈现并讨论单独的比较也许比仅关注于 CR 模型全局化的结果更加符合情理且信息量更多。

CR 模型的目的是在已知已达到或者超过该熟练度水平的情况下，考察超过某一给定熟练度水平的概率。表 5.3 中提供的模型的预测应该理由充分地复制了与各个连续比例相关联的观测条件比例 $P(Y>j \mid Y\geqslant j)$。基于 CR 模型分析的结果，其将性别的作用在连续比例分割点间约束为均等，模型的预测看起来与观测到的条件概率非常接近，对男孩和女孩都是如此。为了今后的参考，表 5.3 也提供了估计的互补概率。这些互补概率是 $P(Y=j \mid Y\geqslant j)$，来自作者之前对风险或者失败率的估计的讨论，也即未能超过熟练度类别 j 的概率。

这些数据的 Somer's D 是 0.691，表明在超过某给定类别的观测概率和预测概率之间良好的对应，尽管这个等级相关系数(rank-order correlation coefficient)是重构的二分结果数据而非原始定序因变量的。似然比 $R_L^2=0.289$，表明 CR 模型相对于虚无模型或截距模型在变异上的适当缩小；再次，该系数是针对重构的二分结果数据集的。

为了通过 τ_p 和 λ_p 考察预测有效性，可以使用估计的条件概率来构建分类表。SAS 将会把 $N=13053$ 个个案的每个 P(“超过” = 1)存储到重构的数据集中。然而，将该数据集汇总到原始规模的样本是有问题的，因为仅有那些包含在各个条件比较的人群才会在那个分割中有预测。因此，生成分类表的简便办法是将模型估计输入统计软件并在原始数据集中使用这些估计 $(N=3365)$ 来计算基于性别的 logit。然后 logit 可以被转化为条件比例，$\hat{\delta}_j=\exp(\text{logit}_j)/$

$(1+\exp(\text{logit}_j))$，于是各个儿童都有一个对每个条件比较 $P(Y>$类别$j \mid Y\geqslant$类别$j)$ 的估计，该过程将得到与表 5.3 中相同的估计。

于是可以使用以下关系式得到类别概率(Läärä & Matthews, 1985)：

$$(1-\delta_j(\underline{x}))=\frac{\pi_j(\underline{x})}{1-\gamma_{j-1}(\underline{x})}$$

这里 $\delta_j(\underline{x})=P(Y>$类别$j \mid Y\geqslant$类别$j,\ \underline{x})$，$\pi_j(\underline{x})=P(Y=$类别$j \mid \underline{x})$，$\gamma_j(\underline{x})=P(Y\leqslant$类别$j \mid \underline{x})=\pi_0(\underline{x})+\pi_1(\underline{x})+\pi_2(\underline{x})+\cdots\pi_j(\underline{x})$。移项以求得类别概率 $\pi_j(\underline{x})$，得到：$\pi_j(\underline{x})=(1-\delta_j(\underline{x}))\times(1-\gamma_{j-1}(\underline{x}))$。这就得到六个类别概率，响应变量的每个水平各一个。个体类别属性被指派到该儿童所拥有的最大概率的类别上。该过程的句法在附录的句法 C8 中可以找到。

对于仅有性别的模型，logistic CR 模型将所有的儿童预测到类别 3。这与第 4 章的 CO 模型类似，$\tau_p=0.23$，$\lambda_p=0$。尽管估计概率的模式可能是有信息的，但单变量模型看起来在熟练度分类精确性上不是很有用。然而，这里所勾画出的方法可以应用到更复杂的 CR 分析以及其他定序数据集中领域中。

使用互补双对数关联的连续比例模型

由于其与隐含的二分 logistic 回归模型的直接对应关系，logit 关联函数经常在需要连续比数模型的研究中使用。一个替代的关联函数是互补双对数(clog-log)关联函数，它以风险比而非比数比的形式提供对结果的解释。附录 C3 中包括了使用重构数据集的互补双对数关联 CR 分析的句法，与之前的 logit 关联分析相比有两处变动。第一，关联要求是互补双对数。第二，使用的是升序方法而非降序方法，为了对应于作为预测风险的(转换)模型估计的解释。即对于重构数据集，我们希望预测的是 P(“超过” $=0$) 而非 P(“超过” $=1$)。

条件概率 $P(Y_i=$类别$j \mid Y_i\geqslant$类别$j)$ 提供了对风险的估计。使用互补双对数关联的升序选项，该估计概率是 logit 关联模型的互补；也即 logit 模型估计的是超过类别 j 的条件概率，或者是 $P(Y_i>$类别$j \mid Y_i\geqslant$类别$j)$，而这里构建

的互补双对数模型估计的是在类别 j 上停止并退出(stopping out)的条件概率 $P(Y_i = \text{类别} j \mid Y_i \geqslant \text{类别} j)$。这种对应允许两种关联函数估计的连续比例和/或它们的互补可以直接比较。不用升序选项时，拟合的互补双对数模型会估计出一套完全不同的条件概率，名义上是 $P(Y_i < \text{类别} j \mid Y_i \leqslant \text{类别} j)$。我们必须谨慎行事以确保合适的兴趣事件(这里是 $Y = \text{类别} j \mid Y \geqslant \text{类别} j$) 以及它的互补(这里是 $Y > \text{类别} j \mid Y \geqslant \text{类别} j$) 与研究者所感兴趣的预测相匹配。表 5.4 中的前两列提供了使用 SAS 的 logistic 回归过程的分析结果。

表 5.4　使用互补双对数关联的 CR 模型参数估计，基于重构的数据集 $N = 13053$，以及原始数据集 $N = 3365$

重构		原始	
Intercept	−0.1166*	Intercept0	−4.0091**
dumcr0	−3.8924**	Intercept1	−2.3257**
dumcr1	−2.4146**	Intercept2	−1.2173**
dumcr2	−1.5013**	Intercept3	0.1444**
dumcr3	−0.0350	Intercept4	0.7155**
性别	**0.1976****	**性别**	**0.1976****
Model fit			
−2LL(model)	10026.201		10026.201
−2LL(null)	14092.374		10053.980
χ^2(df)	4066.173** (5)		27.7511** (1)

注：$*p < 0.05$；$**p < 0.01$。

互补双对数关联函数转换观测概率的方式与使用 logit 关联的方法非常不同。通过互补双对数关联函数得到的值不是比数的对数，而是风险的对数，因此它代表了比例风险模型。与前一部分例 5.1 中的 logit 关联类似，响应变量基于条件概率转换：$p = P(Y = \text{类别} j \mid Y \geqslant \text{类别} j)$。然而，这里 logit 被定义为比数的对数，互补双对数被定义为互补概率负对数的对数：clog-log $= \log(-\log(1-p))$。为了使用互补双对数预测将数据还原到条件概率，使用该过程的倒数：$p\text{-}hat = 1 - \exp(-\exp(\text{clog} - \log))$。因为在比例比数模型中，对数据有一项限制性假设：比例风险。该假设要求风险(风险 $= \exp(\text{clog-log})$)在响应变量的各个不同水平上保持同质比例性。

基于附录中的句法 C3，我们可以从表 5.4 中的结果概要看到模型拟合得比虚无模型好，$\chi_5^2 = 4066.173$，$p < 0.0001$。在该分析中，性别对互补双对数的作用是+0.1976，统计上显著不等于 0（$p < 0.01$）。Singer 和 Willett(2003:424)指出不论选择何种关联函数，我们都将模型中的估计指数化以方便解释："一个来自 logit 关联模型的反对数系数是比数比，而一个来自互补双对数关联模型的反对数系数是风险比。"将当前模型的性别作用指数化，我们得到风险比 HR = exp(0.1976) = 1.218。对于从 0 到 5 的所有响应水平，男孩处于特定类别 j 而非超过它的风险，而女孩被假设为固定在 1.28。为了理解这对数据意味着什么，互补双对数模型的模型估计和预测概率在表 5.5a 中提供。记号 $h\hat{\delta}_j$ 用来表示基于互补双对数关联模型的预测条件概率以区分通过 logit 关联分析估计的条件比例。

表 5.5a　$P(Y > j \mid Y \geq j)$ 的观测比例(δ_j)，预测，估计风险和互补，以及性别模型的估计风险比，互补双对数关联的 CR 分析(使用重构数据集)

	类别					
	0	1	2	3	4	5
男孩						
$\delta_{j(obs)}$	0.9713	0.8997	0.7811	0.3564	0.3710	—
clog-log	−3.8114	−2.3336	−1.4203	0.046	0.081	
$h\hat{\delta}_j$	0.0219	0.0924	0.2147	0.6490	0.6619	
$1-h\hat{\delta}_j$	0.9781	0.9076	0.7853	0.3510	0.3381	
女孩						
$\delta_{j(obs)}$	0.9888	0.9313	0.8241	0.4182	0.3836	
clog-log	−4.009	−2.5312	−1.6179	−0.1516	−0.1166	—
$h\hat{\delta}_j$	0.0180	0.0765	0.1780	0.5766	0.5893	—
$1-h\hat{\delta}_j$	0.9820	0.9235	0.8220	0.4234	0.4107	
HR[a] (est.)	1.218	1.218	1.218	1.218	1.218	—

注：$h\hat{\delta}_j = p\text{-}hat = P(Y = \text{类别}\ j \mid Y \geq \text{类别}\ j)$。

a. 风险 = exp(clog-log)；风险比 HR = (风险(男孩)/(风险(女孩))。

通过拟合模型决定的风险，能够最好地理解男孩对女孩的预测概率。通过升序选项，预测是针对 $h\hat{\delta}_j = P(Y = j \mid Y \geq j)$；这与 logit 关联的 CR 分析得到的预测是互补的。通过互补双对数模型得到的互补概率，$1-h\hat{\delta}_j = P(Y > j \mid$

$Y \geqslant j$)，也在表 5.5a 中予以呈现。

在这些模型里，男孩在性别上的编码是性别 = 1，女孩是性别 = 0。切割点使用的虚拟编码表明，截距对应于 $Y = 4$ 到 $Y = 5$ 的最终比较。为了得到男孩的第一个连续比例（将 $Y = 0$ 与 $Y \geqslant 0$ 比较）的互补双对数估计，我们有（表 5.4）：$\text{clog-log}_{男孩,0} = -0.1166 + (-3.8924) \times \text{dumcr0} + (0.1976) \times 性别 = -3.8114$。对于女孩模型估计对应于 $\text{clog-log}_{女孩,0} = -0.1166 + (-3.8924) \times \text{dumcr0} = -4.009$，将各个互补双对数指数化再取男孩对女孩的比值，得到风险比 $HR = \exp(-3.8114)/\exp(-4.009) = 1.218$。这正是模型在所有的连续比例比较中约束为相等的比例。基于表 5.5a 中互补双对数估计而计算出的对各个连续比例比较的风险比都等于 1.218，含舍入误差。男孩未能超过的风险被假设为比女孩的大，并在条件连续比例分割点间一致。这说明男孩不大可能比女孩取得更高的阅读熟练水平，这与 logistic CR 模型的结果是一致的。

表 5.5b　$P(Y > j \mid Y \geqslant j)$ 的观测比例(δ_j)，预测，估计风险和互补，以及性别模型的估计风险比，互补双对数关联的 CR 分析（使用原始数据集）

	类别					
	0	1	2	3	4	5
男孩						
δ_j	0.9713	0.8997	0.7811	0.3564	0.3710	
clog-log	−3.812	−2.122	−1.020	0.3420	0.9131	—
$c\hat{\delta}_j$	0.0219	0.1123	0.3028	0.7553	0.9173	—
$1-c\hat{\delta}_j$	0.9781	0.8877	0.6972	0.2447	0.0827	
女孩						
δ_j	0.9888	0.9313	0.8241	0.4182	0.3836	
clog-log	−4.009	−2.323	−1.217	0.1444	0.7155	—
$c\hat{\delta}_j$	0.0180	0.0931	0.2562	0.6850	0.8707	—
$1-c\hat{\delta}_j$	0.9820	0.9069	0.7438	0.3150	0.1293	
HR[a] (est.)	1.218	1.218	1.218	1.218	1.218	—

注：$c\hat{\delta}_j$ = 累积概率。
a. 风险 = exp(clog-log)；风险比 HR = (风险(男孩)/(风险(女孩))。

为了考察均等斜率假设，将代表了性别和各个分割之间交互的四个交互项加入到模型中，$-2LL_{int} = 10011.002$（分析未予展示），与 $-2LL_{non\text{-}int} =$

10026.201 比较时，差异是 $\chi_4^2 = 15.20$，超过了 $\chi^2_{4,\,0.005} = 14.86$。这表明单个解释变量性别在不同结果类别间的风险不是平行的。然而，为了演示的目的，我们将继续探讨平行斜率模型参数估计和预测概率的解释。

估计的条件概率用 $\mathrm{h}\hat{\delta}_j$ 表示，可以通过性别模型中预测互补双对数的关联函数的倒数来求得。也就是：$p\text{-}hat = \mathrm{h}\hat{\delta}_j = 1 - \exp(-\exp(\text{clog-log}))$。我们可以将互补预测值 $1 - \mathrm{h}\hat{\delta}_j$ 与表 5.5a 中的第一行的观测连续比例相比较，它们与实际数据间紧密的对应关系就很清楚了。

基于互补双对数关联的预测比那些通过 logit 关联得到的预测更精确吗？对于互补双对数关联，$R_L^2 = 0.289$，Somers' D = 0.691，这与 logit 关联模型是一样的(允许舍入)。两种关联函数看起来都表现得很好。就预测效率来讲，我们可以看到两种关联函数背后相同的预测概率模式；简单模型中仅有性别作为预测变量，男孩和女孩的估计熟练度都是 3 级；因此，$\tau_p = 0.23$ 且 $\lambda_p = 0$。分类表是通过一种与 logistic CR 分析之后类似的过程推导出来的，并对预测互补双对数而非 logit 求解。当生成分类表时互补概率没必要解出来(如果将附录 C8 中的句法修改)，因为对于升序互补双对数模型预测概率已经是 $\mathrm{h}\hat{\delta}_j = P(Y = \text{类别}\,j \mid Y \geq \text{类别}\,j)$。基于预测有效性统计，没有表现出对 logit 或者互补双对数关联的明显偏好。然而，logit 模型的确比互补双对数模型更好地满足了均等斜率假定。

关联的选择和两个互补双对数模型的等价

通常，在 logit 或者互补双对数关联函数下的 CR 模型拟合统计量能够比较，如我们在这里看到的。关联函数之间的选择归结到对某种方法优势的偏好问题，以及对数据如何生成的概念思考。logit 关联的优势在于它以比数和比数比的形式来解释的简洁性。互补双对数关联的优势在于它以风险和风险比的形式解释，并且它的方向与比例风险模型相关。对于这里呈现的分析，“次数”并不是数据如何生成的结构性部分，所以支持 logit 方法的观点可能会更强一些。然而，无论哪种关联函数，特定阶段的交互可以直接包括并被嵌套到模型

中，所以对均等斜率假设的考察可以轻易进行。

然而，在计算上，重构大的数据集以拟合连续比例模型可能会变得较为繁琐（get cumbersome），尤其是当响应变量的水平数目增加时。通过互补双对数关联，可以使用原始数据集，并应用前一章的比例比数模型。如马修斯等人（1985）所展示的，这种方法的结果可以与通过应用重构数据集的互补双对数关联得到的连续比例模型直接比较。两种模型得到相同的变量作用，可以提供以风险比的形式在数据中的直接解释模式。进一步说，通过这种方法，SAS 和 SPSS 都提供对均等斜率假设的检验。然而，在使用重构数据集时，CR 模型提供了条件概率的估计，累积 CR 模型提供的是累积概率的估计。附录中的句法 C4 展示了如何在 SAS 中使用升序选项拟合模型；因此，模型预测的是 $P(Y \leqslant \text{类别}\ j)$。等价的 SPSS 命令在附录 C5 中给出。使用原始数据集的性别作为单个解释变量的 SAS 结果在表 5.4 的第二部分给出。

首先要注意到性别的作用，$b=0.1976(p<0.01)$，在两个互补双对数模型中相等。也就是，以解释变量作用的形式，对有二分结果的重构数据集使用互补双对数关联和在累积比数模型中使用互补双对数关联是等价的。两种方法的参数估计都是以风险的形式解释的。如稍早看到的，男孩未能超过某一给定熟练度水平的风险是女孩风险的 $\exp(0.1976)=1.28$ 倍。然而，两个互补双对数关联模型的截距值并不相等，这通常归咎于两种模型迥异的结构。不应期望截距可比，因为一种模型预测条件概率（重构数据集的互补双对数关联）而另一种预测的是累积概率（原始定序数据的互补双对数关联）。以概率的形式进行的预测不等价。

使用互补双对数关联对重构数据的预测在表 5.5a 中给出；使用互补双对数对原始数据集的累积 CR 模型的预测在表 5.5b 中给出。回忆下预测概率是通过 p-hat $=1-\exp[-\exp(\text{clog-log})]$ 得到的。这里，用 $c\hat{\delta}_j$ 来将之前分析的那些概率与这些概率进行区分。累积 CR 模型的预测可以与表 4.2 中的实际累积概率进行比较。从累积概率的形式看，模型似乎很好地反映了数据。然而，一个变量的模型的拟合证据却很小。该分析的 Somer's D 是 0.079，$R_{\mathrm{L}}^2=0.003$。对于虚拟 R^2 我们得到：$R_{\mathrm{CS}}^2=0.008$，$R_{\mathrm{N}}^2=0.009$，$\tau_{\mathrm{p}}=0.23$，$\lambda_{\mathrm{p}}=0$，因为仅

有性别的累积模型将所有的儿童预测到类别 3 中,不论关联函数为何。

均等斜率假设的检验得到一个 $\chi_4^2 = 14.95$, $p = 0.0048$。表明参数估计在类别间不等价。布兰特(Brant, 1990)发现了计分检验的若干问题,即使不考虑关联函数。大样本倾向于得到小的 p 值,所以基于计分检验的结论可能是有缺陷的;也就是说,拒绝比例比数假设的决定也许并不意味着针对任何或者所有解释变量的类别比较间实际含义的不同。此外,计分检验仅提供了对平行斜率假设缺乏拟合的全局评估,并没有关于违背假设的实质等信息。如在第 4 章中强调的,布兰特建议考察一下伴随着定序分析的隐含二分模型,目的是为了"评估模型的充分度,同时获得关于数据复杂性的额外洞见"(Brant, 1990:1176)。限于篇幅,尽管这里没有考察在互补双对数关联函数下的仅有性别的二分模型,但那样做还是很有道理的,可以增加从全局计分检验中获得的信息量。

连续比例模型的方法选择

如我们在第 4 章看到的,当结果变量是定序时,logit 关联被用来拟合比例比数模型。使用 logit 关联的比例比数模型也许是社会科学研究中各种竞争模型间最常选择的,但是对于定序结果,也有使人信服的理由来考虑用连续比例模型来拟合。如我们在这里看到的,有若干种不同的方法发展 CR 模型。从本质上说,当因变量代表相继阶段的定序进步时,如果变量作用是用条件概率(例如,以达到某特定阶段或者没有为条件)而非累积概率估计的话,则会对影响该进步的因素有更好的理解。由于它与隐含的比例风险模型的关系,互补双对数关联能够提供一个简明分析,该分析能够详尽地为这些变量依次对每个条件二分的作用建模。

logistic CR 回归模型"正好是由考克斯(Cox, 1972)提出的对离散生存分布的比例'logit 风险'模型"(C. Cox, 1998:436),以离散时间提供风险。当使用互补双对数关联时,拟合模型是比例风险模型(Ananth & Kleinbaum, 1997)。关联函数间的选择应该由任一转换下所获得的结果的有用性来做决定。考虑到教育学和社会科学领域的诸多应用研究者都对 logit 模型、比数和比数比越来越

熟悉,有充分的理由期待 logistic CR 方法将会是最适合该领域定序结果研究的方法。与使用累积 CR 互补双对数模型相比,应用重构数据集拟合这些模型而不考虑关联函数,也许是非常明智的策略。重构数据集允许直接生成并纳入交互项,这与第 4 章 PPO 例子所采用的方式完全一样。在解释变量的平行比数假设不成立的情况下,用偏 CR 模型(Cole & Ananth, 2001)拟合也许能提供数据中作用的更好解释。

ECLS-K 数据的全模型连续比例分析

本章的最后一部分将提供 ECLS-K 数据的多变量 logistic CR 模型的例子。分析是由附录中的句法 C6 得出的,使用 SAS 的降序选项。兴趣条件概率是含八个变量的模型的 $P(Y>$类别$j \mid Y \geqslant$类别$j)$。表 5.6 提供了分析的结果,表中还包括了对应的二分 logit 分析的结果以用于比较。表 5.6 的最后一列包括通过 SPSS PLUM 拟合的互补双对数 CR 模型(风险)。为了更好地描述,这里把焦点放在 logistic CR 分析的结果上。

总体来看,男孩(gender;比数比 = 0.668),有任何家庭危险因素特征(famrisk;比数比 = 0.808)的儿童,没有父母或者监护人为他们经常阅读的儿童(noreadbo;比数比 = 0.763),以及仅上半天幼儿园的儿童(halfdayk;比数比 = 0.894)更不大可能超过某一特定熟练度水平。家庭社会经济地位(wksesl;比数比 = 1.78)以及儿童的年龄(plageent;比数比 = 1.05)与有着超过某一给定熟练度水平的更大似然性成正相关。

这些作用中的大部分与相对应的二分模型在规模和模式上都是一致的。然而,也有一些明显的差异。如这里编码的少数族裔(minority),在二分分割间的作用变化剧烈;通过 logistic CR 模型估计的少数族裔的全局作用(比数比 = 0.918,统计上不显著)没有充分捕捉到所有 CR 比较的独特作用。此外,上半天幼儿园项目而非全天项目的作用(比数比 = 0.894,统计显著)在 logistic CR 模型中的作用看起来并没有充分反映出幼儿园一天长度的潜在贡献,而这一点在二分 CR 比较中表现很明显。

表 5.6 使用重构数据集的 CR 全模型(logit 关联),$N = 13053$; $P(Y >$ 类别 $j \mid Y \geq$ 类别 $j)$ 的二分 CR_j (logit 关联)分析;以及 SPSS 的互补双对数 PLUM 分析

	Logistic CR b(se(b)) OR	CR_0 b(se(b)) OR	CR_1 b(se(b)) OR	CR_2 b(se(b)) OR	CR_3 b(se(b)) OR	CR_4 b(se(b)) OR	Clog-log CR b(se(b)) HR
Int1[a]	−3.80(0.45)**	3.53(2.11)	−0.85(1.08)	−1.60(0.80)*	−3.51(0.73)**	−3.71(1.21)**	θ_0 = −1.308(0.339)**
Int2(dumcr0)	4.89(0.15)**						θ_1 = 0.391(0.321)
Int3(dumcr1)	3.33(0.10)**						θ_2 = 1.527(0.319)**
Int4(dumcr2)	2.26(0.09)**						θ_3 = 2.960(0.320)**
Int5(dumcr3)	0.20(0.08)*						θ_4 = 3.588(0.322)**
gender	−0.374(0.052)	−1.039(0.283)	−0.489(0.131)	−0.363(0.095)	−0.338(0.087)	−0.119(0.141)	0.242(0.038)
	0.688**	0.354**	0.613*	0.695*	0.713*	0.888	1.274**
famrisk	−0.2137(0.062)	−0.145(0.295)	−0.291(0.146)	−0.169(0.108)	−0.259(0.106)	−0.021(0.180)	0.150(0.045)
	0.808**	0.865	0.748*	0.844	0.772*	0.979	1.162**
center	0.089(0.062)	0.034(0.284)	−0.086(0.149)	0.188(0.108)	0.072(0.108)	0.200(0.187)	−0.069(0.045)
	1.093	0.966	0.917	1.207	1.075	1.221	0.933
noreadbo	−0.271(0.072)	−0.654(0.272)	−0.283(0.154)	−0.200(0.123)	−0.172(0.133)	−0.335(0.245)	0.185(0.052)
	0.763**	0.520**	0.754	0.819	0.842	0.715	1.203**
minority	−0.085(0.058)	−0.228(0.288)	−0.461(0.142)	−0.324(0.103)	0.250(0.097)	0.158(0.156)	0.030(0.043)
	0.918	0.796	0.631*	0.723*	1.284*	1.171	1.03**
halfdayk	−0.112(0.053)	−0.074(0.261)	0.026(0.133)	−0.136(0.097)	−0.254(0.087)	0.130(0.141)	0.079(0.039)*
	0.894*	0.928	1.027	0.873	0.776*	1.139	1.08*
wksesl	0.575(0.039)	1.005(0.168)	0.632(0.100)	0.582(0.076)	0.451(0.063)	0.600(0.102)	0.402(0.029)
	1.777**	2.71**	1.865*	1.790*	1.570*	1.823*	1.495
plageent	0.049(0.007)	0.024(0.032)	0.061(0.016)	0.050(0.012)	0.048(0.011)	0.041(0.018)	0.033(0.005)
	1.05**	1.025	1.062*	1.051*	1.049*	1.042*	1.034
Model χ^2(df)	4548.76**(12)	82.11**(8)	146.74**(8)	180.62**(8)	123.88**(8)	56.58**(8)	449.99**(8)
Parallel Slopes χ^2(df)							134.431(32)**
H−L χ^2(df)		7.796(8)	7.495(8)	13.869(8)	6.118(8)	9.392(7)	
p		0.454(n.s.)	0.484(n.s.)	0.085(n.s.)	0.634(n.s.)	0.310(n.s.)	

注:对于 logit 关联分析,参数估计是对类别 = 1 的分类变量;类别 = 0 的 SPSS PLUM 分析;

a. 对于 Logistic CR,对应于最后的比较:类别 5 和以上(类别 5),通过虚拟编码。

$*p < 0.05$, $**p < 0.01$。

然而，尽管观察到不规则性，logistic CR 分析的模型拟合统计量表明模型的确提供了数据的良好拟合。该分析的关联测量是 Somer's D = 0.739，τ_p = 0.23 和 λ_p = 0.002。对该 Logistic CR 模型的合理调整包括：把少数族裔和入园半天这两项解释变量的交互项也纳入模型，同时把是否去中心的作用从模型中删除（无论儿童在入幼儿园之前是否参与过中心的看护）。

第 6 章 | 相邻类别模型

相邻类别模型概述

定序因变量变量分析(CO 和 CR 模型都是)的第三种方法是以相邻类别配对的形式同时估计解释变量的作用。相邻类别(AC)模型是多分类结果的一般 logit 模型的特殊形式(Clogg & Shihadeh, 1994; Goodman, 1983; Hosmer & Lemeshow, 2000)。在多分类方法中,我们常在各个响应变量和一个基准类别间进行比较,通常是取最后一个响应类别为基准类别。多分类模型是非约束的,因此解释变量的作用可以在各个具体的比较中有所变动。在 AC 模型中,这些作用在相邻类别比较中被约束为恒等或者同质,反映了我们之前描述的 CO 和 CR 模型的比例性和平行性假设。古德曼(Goodman, 1979)将这种假设称为"统一的关联"(uniform association)。

SAS 通过 CATMOD 过程来估计 AC 模型,尽管对于包括了连续解释变量的模型,该分析并不是最优化的。CATMOD 过程通过加权最小二乘法估计模型,这就要求数据以列联表的形式分好组(Allison, 1999; SAS, 1999)。为了将连续变量纳入模型,克洛格和谢哈德(Clogg & Shihadeh, 1994)采取的策略是将其类别化,但如果连续变量在不同响应层次间是非线性的话,这种方法对模型拟合会有影响。样本中与解释变量的值类似的个体的各个属性的样本规模必须足够大,才能得到 AC 模型间比较的可靠估计(Stokes et al., 2000)。尽管当前为拟合 AC 模型而设计的软件存在缺陷,但在那些相邻类别间的比较具有更强的理论感的研究情境下,这种方法的确提供了 CO 或者 CR 方法的合理的替代。

在 AC 模型中,logit 转换将 $\pi_{i,j}$,也即第 i 个人在第 j 个类别的概率,与在

下一个后续的响应类别的概率，$\pi_{i,\ j+1}$进行比较。这种方法的目的是同时决定下一个最高响应的类别对各个相邻类别配对的比数。构建一系列的 logit 比较，对应于这些比较的分析在表 4.1 的最后一部分给出。我们取两个相邻概率的对数来产生 logit：

$$Y_j' = \ln\left(\frac{\pi_{i,\ j+1}}{\pi_{i,\ j}}\right) = \alpha_j + \beta X_i$$

这里 $J =$ 响应水平的个数且 $j = 1, \cdots J-1$。在该式中，由分子和分母呈现的概率是指两个相比较的相邻类别的概率。$J-1$ logit 公式的截距在不同的相邻类别比较间也许会变化，但是解释变量的作用被假设为在不同比较间是一致的。斜率作用 β，在所有的 AC 模型比较间都是同质性的。

AC 模型的比数比被称为"局部比数比"(local odds ratios)，因为他们所描述的关联是来自于总体表格的局部化部分(Agresti，1989，1996)。对于一个有两个水平 0，1 的单个解释变量，$J-1$ 比数比在各个具体的相邻类别对中生成($j = 1$ 到 $J-1$)。

$$OR_{j+1} = \frac{P(Y = j+1 \mid x = 1)/P(Y_i = j \mid x = 1)}{P(Y = j+1 \mid x = 0)/P(Y_i = j \mid x = 0)}$$

对 AC 模型的平行性的考察大部分与之前的分析处理方式相同，但没有对约束的正式检验。如果 AC 模型不拟合，则可以用多分类方法；或者特定 logistic 回归对应于各个 AC 比较的分割可以分别查看约束的可信性。

与之前的章节类似，在接下来的讨论中，仅有性别的模型首先模拟以描述简单 AC 模型的方法、路径和结果。下一步，用更复杂的模型来拟合并解释结果。两个模型的句法在附录的 D 部分呈现(简单的和复杂的)。

仅有性别的模型

在附录 D1 部分呈现的句法，被用于仅有性别的 AC 模型。在该句法以及更复杂的模型里也是，解释变量的作用被视为定量的，因此被包括在 CATMOD 过程里的"直接"(direct)子命令下。因为性别的编码对于 SAS 来说是外部的

(0 = 女性，1 = 男性)，这种方法允许考虑到解释效果的便利性，将解释与之前的那些 logit 模型保持一致。参见斯托克斯(Stokes)等人(2000)对 CATMOD 过程里类别预测变量处理的替代但是等价的方法。“响应”(response)语句指明了相邻类别模型(alogit)，而且句法要求预测的 logit 和估计的 AC 概率写到名为“acgender”的文件里。最终，模型使用一个称做“_response_”的项放在等式右边以具体化。该项将性别的普遍作用约束加在 AC 响应函数间。估计过程细化为加权最小二乘法，也要求一张包含了模型预测 logits 的表格(pred)。

早期阅读的结果有六个可能的响应(0, 1, 2, 3, 4, 5)；因此，他们有 $6-1=5$ 个代表相邻类别比较的响应函数。图 6.1 提供了通过 CATMOD 过程实现的统计分析的主要结果。首先，“总体概要”(population profiles)报告了各个独特共变性模式的样本规模。当只有性别作为预测变量纳入模型中时，样本规模在这里指各个性别群体的样本规模，本表可以用来界定特殊共变模式的稀少样本。

“方差分析”部分提供了性别在 $J-1=5$ 个 AC 响应模型间的性别的作用检验。对应于这个分析，性别的作用是显著的，$\chi_1^2 = 38.38$，$p < 0.0001$。残余卡方检验表明模型的拟合，$\chi_4^2 = 2.96$，$p = 0.5647$。这个卡方检验统计量是一个类似于皮尔森卡方检验的拟合优度检验，是将拟合模型与一个预测模型或者饱和模型进行比较。

加权最小二乘法估计在下一部分的输出结果里呈现。每个相继响应函数的截距通过将各个对应的_response_值与截距项相加得到。表 6.1 描述了 $J-1=5$ 个响应函数的截距的过程。基于表 6.1 展现的结果，这个简单分析的五个响应模型分别是：

$$\hat{Y}'(1, 0) = 1.5725 + (-0.1928) \times \text{性别}$$

$$\hat{Y}'(2, 1) = 0.8688 + (-0.1928) \times \text{性别}$$

$$\hat{Y}'(3, 2) = 1.0146 + (-0.1928) \times \text{性别}$$

$$\hat{Y}'(4, 3) = -0.8347 + (-0.1928) \times \text{性别}$$

$$\hat{Y}'(5, 4) = -0.4147 + (-0.1928) \times \text{性别}$$

Population Profiles

Sample	GENDER	Sample Size
1	0	1692
2	1	1673

Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	176.21	<0.0001
RESPONSE	4	1550.56	<0.0001
GENDER	1	38.38	<0.0001

Analysis of Weighted Least Squares Estimates

Parameter		Estimate	Error	Chi-Square	Pr > ChiSq
Intercept		0.4323	0.0326	176.21	<0.0001
RESPONSE	1	1.0952	0.1158	89.48	<0.0001
	2	0.4365	0.0774	31.81	<0.0001
	3	0.5823	0.0554	110.40	<0.0001
	4	−1.2670	0.0558	515.67	<0.0001
GENDER		−0.1928	0.0311	38.38	<0.0001

Predicted Values for Response Functions

		——Observed——		——Predicted——		
GENDER	Function Number	Function	Standard Error	Function	Standard Error	Residual
0	1	1.800493	0.247643	1.527445	0.139517	0.273048
	2	0.868196	0.111109	0.868796	0.075314	−0.0006
	3	1.002937	0.070628	1.014531	0.051393	−0.01159
	4	−0.81395	0.066029	−0.83477	0.050661	0.020819
	5	−0.47424	0.088744	−0.41475	0.068319	−0.0595
1	1	1.222549	0.16422	1.334673	0.136174	−0.11212
	2	0.674571	0.096229	0.676024	0.073575	−0.00145
	3	0.83155	0.066974	0.82176	0.050663	0.00979
	4	−1.05469	0.072573	−1.02754	0.051874	−0.02716
	5	−0.5279	0.10261	−0.60752	0.069586	0.079621

图 6.1 简单"性别"模型的 PROC CATMOD 结果

表 6.1 $J-1=5$ 个 AC 响应函数的截距

AC 截距	算　式	结　果
α_1	0.4323＋1.0592	1.5275
α_2	0.4323＋0.4365	0.8688
α_3	0.4323＋0.5823	1.0146
α_4	0.4323＋(－1.2670)	－0.8347
α_5	0.4323－[1.0952＋0.4365＋0.5823＋(－1.2670)]	－0.4147

在这些表达式中，等式左边括号里的值指出相比较的两个相邻类别。这些模型可以用来发现对应于各个响应函数的依靠于性别的预测 logit；预测的 logit 再可以转化以估计暗示的两个相邻变量中处于更高类别的概率的比数。估计的响应变量 logit 在输出结果的最后一部分给出“响应函数的预测值”。观察的 logit 在表的前半部分呈现，例如，为了计算男性(性别 = 1)在第四个 AC 比较中的预测 logit(公式 4 代表类别 4 与类别 3 相比)，我们使用以上的关联响应模型：$Y(4,3)=-0.8347+(-0.1982)\times$性别$=-0.8347+(-0.1982)=-1.0275$。我们可以使用同样的过程来得到女孩(性别 = 0)的对应预测 logit：$Y=-0.8347+(-0.1982)\times$性别$=-0.8347$。如果预测 logit 经过指数化，结果就是对于各个性别处在类别 4 而非类别 3 的比数。对于男孩，$\exp(-1.0275)=0.3579$，对于女孩则是 $\exp(-0.8347)=0.4340$。男孩和女孩都不大可能处于类别 4，而非类别 3，这与 CO 和 CR 模型的结果类似。为了计算男孩对女孩的比数比，我们得到 $0.3579/0.4340=0.8246$，即 AC 模型的性别作用的指数值，也等于 $\exp(-0.1982)=0.8246$。对于所有的五个 AC 响应函数，性别的作用以一个对应的共同为 0.8246 的比数比保持一致。因此，根据模型，男孩处于两个相邻类别中较高者的比数是女孩比数的 0.8246 倍；男孩更不大可能处于较高的阅读熟练度水平。

熟练度的分性别的实际频次和类别概率稍早在第 4 章的表 4.2 中(表格第一二行的男性和女性部分)给出。AC 模型并不预测类别概率；确切地说，从 AClogit 模型中得到的预测值可以用来估计两个相比较的相邻类别中更高的响应变量的条件概率。然而，预测的 logit 可以转化为比数，再将比数以通常的方

式转化成预测概率：*p-hat* = 比数 /(1 + 比数)。

通过使用表 4.2 中的数据，以及 AC 模型的结果，计算出观测的和预测的 AC 概率并在表 6.2 中呈现出来。观测的和估计的比数比也在表 6.2 中呈现。概率是以比较的具体相邻类别为条件的。例如，男性的第一个条目 0.7725 代表了男性处于熟练度 1 级而非(下一个稍低的)相邻类别，熟练度 0 级的观测概率。该值是由表 4.2 中这两个相邻类别的条目所决定的：0.7725 = 163/ (48 + 163)。通过使用表 4.2 中具体类别概率也可以得到相同的值：0.7725 = 0.0974/(0.0287 + 0.0974)。

从 AC 模型中得出的预测条件概率在表 6.2 的底部给出。为了从 logit 模型中得到预测概率，我们使用类似的表达式：$\hat{p} = \exp(\log it)/(1+\exp(\log it))$。对于处于比较(4, 3)中的男孩，这就成为 $\hat{p}_{4,男性} = 0.3579/(1+0.3579) = 0.2636$，如表 6.2 中底部的第四列所示。类似的，我们可以得到女孩的 $\hat{p}_{4,女性} = 0.3026$。这些概率是条件的，对应的含义是在给定两个相邻类别之中任一条件下它们所对应的处于更高类别的概率。最后，从观测到估计比数比的直观比较来说，性别作用在 AC 类别间的单个值看起来提供了数据的粗略描述。

总体来看，模型表明性别有助于解释相邻熟练度类别间概率的差异。仅有性别的模型是对没有拟合数据的虚无模型的改进：虚无模型的残余卡方是 $\chi^2_5 = 41.33$，$p < 0.0001$ (句法 D2，结果未予呈现)。基于性别模型，我们可以以下结论说明男孩跟女孩相比更不大可能处于两个相邻类别中较高的阅读熟练度。

表 6.2 观测到的(π_j^*)和预测的($\hat{p}_j^*$)条件 AC 概率

AC Comparison	(1, 0)	(2, 1)	(3, 2)	(4, 3)	(5, 4)
$\pi^*_{j,男性}$	0.7725	0.6625	0.6967	0.2580	0.3710
$\pi^*_{j,女性}$	0.8582	0.7043	0.7316	0.3071	0.3836
OR(observed)	0.5588	0.8245	0.8428	0.7858	0.9474
$\hat{p}^*_{j,男性}$	0.7916	0.6629	0.6946	0.2636	0.3526
$\hat{p}^*_{j,女性}$	0.8216	0.7045	0.7339	0.3026	0.3978
OR(estimated)	0.8246	0.8246	0.8246	0.8246	0.8246

因为从模型中估计的概率是条件的，又因为每一个的比较基础都不一样，

所以并没有估计 AC 模型的特定类别概率的直接方法。测量关联的统计量例如 τ_p 和 λ_p 用这些类别概率来构建分类表。阿格雷斯提(Agresti, 1989)提供了计算类别频次的句法和例子,该计算是基于根据基线类别(多分类)logit 模型的调整和将解释变量的斜率约束为在 AC 响应函数间保持一致的特定设计的矩阵结构。这种方法在解释变量增加时将会变得十分复杂,这里不再叙述。然而,在 AC 模型中的实际概率和预测概率之间有很强的关联,皮尔森的伽马系数为 0.997。这个相关是为两个剖面(性别)在 5 个响应模型中而计算的,因此得到这么高的关联系数并不奇怪。如稍早提到的,CATMOD 使用加权最小二乘法而非极大似然值法来估计 AC 模型;因此,并不存在一个将之前两个定序回归方法比较的似然比 R^2。

两个解释变量的相邻类别模型

CATMOD 过程并不是用来处理连续解释变量的,而且独特共变模式在数据集内的样本量很小时估计会出问题。之前讨论的全模型 AC 与 CO 和 CR 模型的等价性由于这个限制并不可估。AC 模型可以通过 GENMOD 过程或者 SPSS 的 GENLOG 和来自于无约束的多分类模型的调整来拟合,但是那些方法与本书呈现的回归方法多少有些不同,因此在这里不予考虑。多分类结果的分析细节可以在阿格雷斯提(Agresti, 1990, 1996), Allison(1999), Borooah (2002), Ishii-Kuntz(1994),以及坦巴克尼克和芬戴尔(Tabachnick & Fidell, 2001)等人的研究中找到。

为了描述稍复杂的设计 AC 分析的回归方法,本书选取了一个连续变量,上幼儿园的年龄(plageent),构建了一个二变量的模型(入园年龄和性别)。按照克洛格和谢哈德(Clogg & Shihadeh, 1994)使用的有连续预测变量的 AC 模型方法,将年龄划分成 4 层(年龄类别,agecat),开始于 57 个月(4.75 年),各个跨度约为 6 个月。各个解释变量(性别和年龄类别)都通过子命令"直接"(句法 D3)来作为定量的处理。

分析产生了八个共变剖面,两个解释变量的 2 × 4 个交互分类各一个。剖

面内部的样本规模从 18 到 850 不等。残余卡方检验表明模型拟合良好，$\chi^2_{33} = 30.49$，$p = 0.5929$，分析中稍大一点的 p 值表明，相对仅有性别的模型有稍微的改进。性别（χ^2_1 的 $= 42.06$，$p < 0.0001$）和年龄类别（χ^2_1 的 $= 60.43$，$p < 0.0001$）在模型中都统计显著。

五个 AC 模型在下面显示。各个模型的截距形成遵循表 6.1 中勾勒的相同模式。性别作用的比数比是 $\exp(-0.2043) = 0.8152$，这与之前它在仅有性别的分析中的作用类似。因此，控制了入园年龄后，男孩更不大可能比女孩处于两个相邻熟练度类别中较高的熟练度水平。年龄对熟练度的比数比是 $\exp(0.1607) = 1.1743$。控制了性别，年长的儿童与年幼儿童相比更有可能在阅读方面处于较高的熟练度类别。

$$\hat{Y}'(1, 0) = 1.1373 + (-0.2043) \times \text{性别} + (0.1607) \times \text{年龄类别}$$

$$\hat{Y}'(2, 1) = 0.5544 + (-0.2043) \times \text{性别} + (0.1607) \times \text{年龄类别}$$

$$\hat{Y}'(3, 2) = 0.6871 + (-0.2043) \times \text{性别} + (0.1607) \times \text{年龄类别}$$

$$\hat{Y}'(4, 3) = -1.1718 + (-0.2043) \times \text{性别} + (0.1607) \times \text{年龄类别}$$

$$\hat{Y}'(5, 4) = -0.767 + (-0.2043) \times \text{性别} + (0.1607) \times \text{年龄类别}$$

性别和年龄分类的平行性假设是通过回顾对应于各个响应函数潜在的 AC 模型来验证的（分析未予展示）。对于两个解释变量，考虑到各个模型推出的 logits 和比数比，这个解释似乎是合理的。八个剖面（profiles）在五个响应函数间的观测值和预测值之间的关联很强，$r = 0.903$。

全相邻类别模型分析

为了提供前两章结果之间的比较，对应的 AC 二分 logistic 回归是基于整套预测变量的。表 6.3 中展示了结果。粗体印刷的比数比统计上显著不同于 1。在作用的方向上，结果与 CO 模型和 CR 模型的发现完全一致。通常，年长的和来自较高社会经济地位家庭的儿童倾向于处于更高的阅读熟练类别。男孩如

表 6.3 全模型的相邻分类二分 logits

Comparison	(1，0) b[se(b)] OR	(2，1) b[se(b)] OR	(3，2) b[se(b)] OR	(4，3) b[se(b)] OR	(5，4) b[se(b)] OR
Intercept	4.10(2.15)	−0.34(1.20)	−1.18(0.84)	−2.85** (0.84)	−3.71** (1.21)
gender	−0.64(0.31) 0.53*	−0.23(0.15) 0.80	−0.26(0.10) 0.77**	−0.29(0.10) 0.75**	−0.12(0.14) 0.89
famrisk	0.19(0.33) 1.21	−0.18(0.17) 0.83	−0.09(0.11) 0.91	−0.26(0.12) 0.77*	−0.02(0.18) 0.98
center	0.05(0.32) 1.06	−0.24(0.17) 0.78	0.18(0.11) 1.20	0.00(0.12) 1.00	0.20(0.19) 1.22
noreadbo	−0.36(0.30) 0.70	−0.13(0.18) 0.88	−0.14(0.13) 0.87	−0.07(0.15) 0.93	−0.34(0.25) 0.72
minority	0.28(0.33) 1.32	−0.18(0.17) 0.84	−0.41(0.11) 0.67**	0.17(0.11) 1.19	0.16(0.16) 1.17
halfdayK	−0.07(0.30) 0.94	0.10(0.15) 1.11	−0.07(0.10) 0.94	−0.30(0.10) 0.74*	0.13(0.14) 1.14
wksesl	0.92(0.27) 2.51**	0.24(0.12) 1.27	0.40(0.08) 1.50**	0.25(0.07) 1.28**	0.60(0.10) 1.82**
plageent	−0.03(0.03) 0.97	0.03(0.02) 1.03	0.036(0.01) 1.04**	0.03(0.01) 1.03**	0.04(0.02) 1.04*
Model $\chi^2(df)$	22.75** (8)	20.17** (8)	97.61** (8)	44.27** (8)	56.58** (8)
H−L $\chi^2(df)$	7.36(8)	7.25(8)	12.70(8)	5.40(8)	9.40(8)

注：* $p<0.05$；** $p<0.01$。

果来自存在所谓的风险因素的家庭(见第 2 章)，另外，在某种程度上，仅上半天幼儿园而非全天幼儿园的，都倾向于处于较低的熟练度类别。没有经常给他们读书的儿童也倾向于处于较低的熟练度类别，该变量的比数比也通常小于 1，尽管该作用在 AC 模型中统计上都不显著。

第 7 章 | 结论

本书的目的是描述定序因变量变量的分析统计技术，并使得实证研究者熟悉定序数据的分析方法，这些数据是忠实于结果测量实际水平的。使用来自NCES 的儿童早期追踪研究—幼儿园同期群(ECLS-K)的数据，可以用三种定序回归方法予以呈现：比例或者累积比数、连续比数以及相邻类别模型。此外，这些模型的一些扩展也得以呈现，允许对一些解释变量比例性或者平行性假设的宽松约束。这里描述的这些方法和例子使得研究者在他们的研究数据结果以定序因变量形式出现时可以使用类似的模型。

定序因变量变量的分析方法要求一个由研究问题而非“拟合”特定模型的需要来指引的考虑周到和敏感的策略。面对诸多分析的选择，这里提到的是最常使用的定序回归方法。以上界定的模型在模型预测和解释变量的作用上各不相同；因此，模型方法的选择应当一直由理论指引，或者是解释变量如何可能影响定序结果，或者是定序分数如何推导出来。选择还应当由特定的研究目的，以及特定统计模型应用而得到的结果的预期含义加以引导。

克利夫(Cliff, 1993, 1994, 1996b; Cliff & Keats, 2003)等人一直是需要将定序变量作为定序看待的坚定支持者，也就是通过称呼并保存那些变量实际形成的过程来考虑定序数据的分析。他的研究指引了本书大部分关于使用和分析定序结果数据的重要思考。然而，本书的目的不是倡导一种附会于简单的基于响应变量尺度的方法论。相反，目标是在定序分数的解释十分重要时，促进定序技术的理解和使用。

对定序的掌握类型分数的分析，例如从 ECLS-K 数据中得到的早期读写熟练度的尺度，试图理解为何一些儿童能够成功达到某个早期阅读技能，而其他

的儿童未能达到同样的技能，而且它也有助于研究者根据儿童所处的熟练度连续体找出或者发展出意欲改善个体熟练度的干预方法。试图将儿童从最低的熟练度类别上升到任何更高的熟练度的相同干预也许并不适用于已经处于两三个最高类别上的儿童。定序分数分析在其他领域也有类似的用处。例如，减小风险的干预，那些意欲促进安全套使用的措施，对于不同打算改变阶段的目标人群经常使用不同的干预信息（Prochaska, DiClemente et al., 1992; Stark et al., 1998）。因此，一个尺寸并不能适合所有定序结果变量的内容。

在所呈现的例子中有很多作用方向的类似性。例如，那些有家庭风险或者没有家长或者监护人为他们读书的一年级孩子更不可能比他们的同龄人处于更高的熟练度水平。年长的或者来自较高社会经济地位的一年级儿童相比他们的同龄人更有可能处于较高的熟练度水平。观察全 CO 模型和 CR 分析的表格，所有作用的比数比在方向上都是类似的，但是比数比的解释要取决于所构建的特定模型。累积比数是用来呈现儿童处于或者高于任意特定熟练度类别的比数。连续比例是在已知该儿童已经达到该熟练度类别的情况下，用来呈现儿童超过某特定类别的比数。最后，AC 模型是被设计用来估计儿童处于两个相邻熟练度水平中较高者的比数。一个最重要的问题是："这些模型中的哪一个是'最好的'？"

答案很简单：这取决于研究问题。CO 模型在研究兴趣是澄清结果的趋势时尤其有用，无论解释变量的不同值是向上还是向下的（Agresti, 1996）。CR 模型可能在发展类型的研究中最有用，研究者的最大兴趣是意欲弄清楚在给定某特定阶段已经达到的情况下（O'Connell, 2000），与在响应连续体上更进一步关联的因素。AC 模型宣称哪一个预测变量最好地预测了处于下一个最高响应类别的响应，因此有助于找出在 AC 响应配对之间的差异。

然而，"最好"的定序模型也将是比例或者平行性假设合理的那一个。如这里所做的一样，拟合并观察各个方法相对应的二分模型，补充一些对这些假设的检验 CO 和 CR 模型能够调整为允许隐含的分割点或者类别区分间的交互项；这就是偏比例比数模型，也可以成为偏比例风险模型。尽管这两种交互模型都要求重构数据集，他们在理解解释变量对响应连续体变动的不同作用上仍

然非常有用。对于 ECLS-K 研究，偏比例风险模型也许最好地呈现了熟练度过程以及与一个儿童实际上能够取得更高类别熟练度的可能性相关联的因素。

尽管来自 ECLS-K 数据的定序熟练度分数分析的目标也许支持了作者在这个特殊的例子中对连续比例模型的偏好，但此处并没有尝试将一个定序回归方法或多或少地置于比其他替代方法更合适的位置。每个模型都将不同的假设集置于数据之上，并引出不同种类的研究问题。研究者的责任就是完全理解这些假设的本质以及如何做出应用一种模型而非另一种也许会影响研究发现的有用性。

假设数据的测量属性比实际的更强，例如将定序结果的值看成是均等间距尺度，则模糊了数据模式的丰富性，这种威胁通过专为定序结果设计的方法将得到更好的暴露。第 4 章包括了累积比数模型和多元线性回归模型结果的比较，这种直接比较的结果应该传达给研究者这样一种信号：依靠熟知的方法，例如多元回归分析，会遮盖重要作用的解释并且导致毫无道理的预测。另一方面，完全忽视数据的定序性，即将结果作为严格的名义变量对待，阻碍了评估方向性和进步的能力，而这看起来正是一开始之所以构建定序因变量测量的关键。我们希望，研究者们通过使用本书中所呈现的研究，能在他们的分析方法中考虑定序回归技术，尤其是当他们的问题和数据表明这样的方法正确的时候。

进一步研究的考虑

在很多方面，使用定序回归模型仍然是一个发展中的方法。针对那些希望今后能够发展并应用定序模型的读者，作者提出以下兴趣点：

- 这里呈现的关联的测量，包括 τ_p 和 λ_p，以及似然比 R^2 统计量，是很微弱的。这也许是数据的加工品(artifact)；熟练度的分布是不均衡的，很大一部分(44%)的儿童在一年级一开始就已取得了熟练度为 3 级的水平。朗(Long, 1997)指出，尽管有必要对概要的统计进行拟合，但“在实际中它们是有问题的”，因为它们只提供了模型质量的部分信息。霍斯默和莱默

苏(2000)提出了有趣的观点:从 logistic 回归模型中得出的分类统计之所以不合适,是因为分类倾向于被样本分布的概率所左右。尽管他们的论断是特指二分模型,扩展到定序模型的分类,类似的问题也很明显。无论如何,在不同样本中考察定序因变量的关联的不同测量的行为能够为这个领域提供很有价值的信息。

- 考察定序方法的比例性或者平行性假设的做图方法,以及残差诊断方法都没有得到很好的发展。SAS 或者 SPSS 都未能在它们的程序中包括定序回归的残差诊断(尽管二者都有对 logistic 回归的)。尽管这里没有考察图形和残差的诊断,本德和本纳(Bender & Benner, 2000)展示了他们所考察的定序模型子集的若干诊断和做图方法。
- 这里呈现的方法假设个体间的独立,是用个体水平的模型拟合数据。然而,ECLS-K 研究是基于多阶段分层抽样策略,从抽中的学校中再抽取儿童。因为本研究的目的是详细解说定序回归模型的应用,就没有包括数据的多层结构调整。定序数据的多层模型可以通过主要的多层软件包来实现,它们包括:HLM (Raudenbush, Byrk, Cheong, & Congdon, 2000), MLwiN(Goldstein et al., 1998),以及 MIXOR(Hedeker & Gibbons, 1996)。HLM 能够拟合多层数据的比例比数模型。MLwiN 能够拟合比例比数模型,并可以通过重构数据估计多层 Logistic 连续比例模型以及 Logistic 偏或者非比例风险模型。MIXOR 是专为定序结果的多层分析而设计的,它比 HLM 或者 MLwiN 的灵活性都要大。关联函数包括 logit, probit 以及互补双对数。

附录一 | 第 3 章

“Gonomiss”是兴趣解释变量无缺失观测的子样本 SAS 数据集。“ECLS-FGsub”是兴趣解释变量无缺失观测的子样本 SPSS 数据集。

A1. SPSS LOGISTIC(对于熟练度结果 0，1 VS. 5)

```
** (cumsp2 = 0 if profread = 0, 1; else cumsp2 = 1 if profread = 5).
temporary.
select if(profreadle 1 OR profread eq 5).
logistic regression CUMSP2
     with GENDER
     /print = all
     /save = pred.
```

A2. SAS PROC LOGISTIC(降序选项;其他选项如下)

```
** data "go" contains only children with values of 0, 1, or 5 on profread;
** cumsp2 = 0 if profread = 0, 1_else cumsp2 = 1 if profread = 5;
proc logistic data = go order = internal descending;
     model cumsp2 = gender /link = logit lackfit ctable pprob = .5001 rsquare;
     output out = dataprobs pred = phat;
run;
```

A3. SAS PROC LOGISTIC(默认为升序选项;基本选项)

```
** 数据“go”仅包括“profread”的值为 0, 1 或者 5 的儿童;
** cumsp2 = 0 if profread = 0, 1_else cumsp2 = 1 if profread = 5;
proc logistic data = go order = internal;
     model cumsp2 = gender /link = logit rsquare;
run;
```

A4. SPSS PLUM

```
temporary.
select if(profreadle 1 OR profread eq 5).
PLUM
cumsp2 BY gender
/LINK = logit
/PRINT = FIT PARAMETER SUMMARY TPARALLEL HISTORY(1) KERNEL
/SAVE = ESTPROB PREDCAT PCPROB ACPROB.
```

附录二 | 第 4 章

B1. SAS 升序(CO 模型),X1=性别(Gender)

```
proc logistic data = sagebook.gonomiss;
    model profread = gender /link = logit rsquare;
    output out = propodds predprobs = cumulative;
run;
```

B2. SAS 降序选项(CO 模型),X1=性别(Gender)

```
proc logistic data = gonomiss descending;
    model profread = gender /rsquare;
    output out = proppred predprobs = cumulative;
run;
```

B3. SPSS PLUM(CO 模型),X1=性别(Gender)

```
通过 filt_$ $ 筛选.    ** 筛选出含缺失值的个案**
    PLUM
        profread BY gender
        /LINK = LOGIT
        /PRINT = FIT PARAMETER SUMMARY TPARALLEL HISTORY(1) KERNEL
        /SAVE = ESTPROB PREDCAT PCPROB ACPROB.
```

B4. SAS 全模型累积比数模型(降序)

```
proc logistic data = sagebook.gonomiss descending;
    model profread = gender famrisk center noreadbo minority halfdayK wksesl plageent
    /link = logit rsquare;
    output out = proppred predprobs = cumulative;
run;
```

B5. 通过 SAS 的 GENMOD 实现的偏比例比数

```
**** 为了偏比例比数生成重构的数据集*** ;
data ppom; set gonomiss;
    do; if profread=5 then beyond=1;
    else beyond=0; split=5; output; end;
    do; if profread ge 4 then beyond=1;
    else beyond=0; split=4; output; end;
    do; if profread ge 3 then beyond  =1;
    else beyond=0; split=3; output; end;
    do; if profread ge 2 then beyond=1;
    else beyond=0; split=2; output; end;
    do; if profread ge 1 then beyond=1;
    else beyond=0; split=1; output; end;
run;

proc freq data=ppom;
    tables split* profread* beyond;
run;

proc sort data=ppom;
    by split gender famrisk center noreadbo minority halfdayK;
run;

    ** 使用与单个的 logit 分析最类似的 INDEP 结构;
    ** 参见 Stokes, Davis, Koch(2000),第 541 页;

proc genmod descending order=data data=ppom;
    class split gender famrisk center noreadbo minority halfdayK childid;
    model beyond=gender famrisk center noreadbo minority halfdayK wksesl plag-
eent split split* minority
    /link=logit d=b type3;
    repeated subject=childid /type=indep;
run;
```

附录三 | 第 5 章

C1. SAS:重构数据集并对 P(超过)建模

```
data cr1;
    set gonomiss;
    if profread ge 0; crcp = 0;                    ** 连续比例切割点** ;
    beyond = profread ge 1;                        ** 否则 = 0** ;
run;

data cr2;
    set gonomiss;
    if profread ge 1; crcp = 1;
    beyond = profread ge 2;
run;

data cr3;
    set gonomiss;
    if profread ge 2; crcp = 2;
    beyond = profread ge 3;
run;

data cr4;
    set gonomiss;
    if profread ge 3; crcp = 3;
    beyond = profread ge 4;

data cr5;
    set gonomiss;
    if profread ge 4; crcp = 4;
    beyond = profread ge 5;
run;

data concat;
    set cr1 cr2 cr3 cr4 cr5;
    if crcp = 0 then dumcr0 = 1; else dumcr0 = 0;
    if crcp = 1 then dumcr1 = 1; else dumcr1 = 0;
    if crcp = 2 then dumcr2 = 1; else dumcr2 = 0;
    if crcp = 3 then dumcr3 = 1; else dumcr3 = 0;
run;
```

C2. SAS：logit 关联连续比例(CR) 模型，使用降序选项

```
proc logistic data = concat descending;
    model beyond = dumcr0 dumcr1 dumcr2 dumcr3 gender /link = logit rsquare;
    output out = modC2 pred = phat;
run;
```

C3. SAS：互补双对数关联连续比例(CR)模型，使用升序(默认)选项

```
proc logistic data = concat;
    model beyond = dumcr0 dumcr1 dumcr2 dumcr3 gender /link = cloglog rsquare;
    output out = modC3 pred = phat;
run;
```

C4. SAS：互补双对数关联累积连续比例(CR)模型，使用升序(默认)选项

```
proc logistic data = gonomiss;
    model profread = gender /link = cloglog rsquare;
    output out = modC4 pred = phat;
run;
```

C5. SPSS：互补双对数关联累积连续比例模型

```
PLUM
    profread BY male
    /CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100)
    MXSTEP(5)
    PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
    /LINK = cloglog
    /PRINT = FIT PARAMETER SUMMARY TPARALLEL HISTORY(1) KERNEL
    /SAVE = ESTPROB PREDCAT PCPROB ACPROB.
```

C6. SAS：logistic 连续比例，全模型

```
** CR1 logit link;
proc logistic data = sagebook.concat descending;
    model beyond = dumcr0 dumcr1 dumcr2 dumcr3 gender famrisk center noreadbo mi-
        nority halfdayK wksesl plageent
    /link = logit rsquare;
    output out = modC6 pred = phat;
run;
```

C7. SPSS：累积连续比例，全模型

```
PLUM
    profread BY male famrisk center noreadbo minority halfdayk
    WITH
    plageent wksesl
    /CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
    PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
    /LINK = cloglog
    /PRINT = FIT PARAMETER SUMMARY TPARALLEL HISTORY(1) KERNEL
    /SAVE = ESTPROB PREDCAT(CRpred) PCPROB ACPROB.
```

C8. 在 SPSS 中生成基于从句法 C2(SAS 降序)得到的模型估计的 logistic 连续比例分析的分类表

```
* CR logit 模型 P(超过 = 1) ***********
* 使用定序数据集,n = 3365 ***********
compute int = -.3763.
compute dumcr0 = 4.4248.
compute dumcr1 = 2.9113.
compute dumcr2 = 1.9283.
compute dumcr3 = 0.0578.
compute slopegen = -0.2865.
compute logit4 = int + slopegen* gender.
compute logit3 = int + dumcr3 + slopegen* gender.
compute logit2 = int + dumcr2 + slopegen* gender.
compute logit1 = int + dumcr1 + slopegen* gender.
compute logit0 = int + dumcr0 + slopegen* gender.
```

```
compute delta0 = exp(logit0) /(1 + exp(logit0)).
compute delta1 = exp(logit1) /(1 + exp(logit1)).
compute delta2 = exp(logit2) /(1 + exp(logit2)).
compute delta3 = exp(logit3) /(1 + exp(logit3)).
compute delta4 = exp(logit4} /(1 + exp(logit4)).
* freq /var = delta0 delta1 delta2 delta3 delta4.
***** 现在需要德尔塔的互补以用于降序 logit 关联****
compute compd0 = 1 - delta0.
compute compd1 = 1 - delta1.
compute compd2 = 1 - delta2.
compute compd3 = 1 - delta3.
compute compd4 = 1 - delta4.
compute compd5 = 1.0.
* freq /var = compd0 to compd5.
* freq /var = p0 to p5.
compute p0 = compd0.
compute p1 = compd1* (1 - p0).
compute p2 = compd2* (1 - p0 - p1).
compute p3 = compd3* (1 - p0 - p1 - p2).
compute p4 = compd4* (1 - p0 - p1 - p2 - p3).
compute p5 = compd5* (1 - p0 - p1 - p2 - p3 - p4).
***** 现在,将最大的类别概率作为选择预测的类别值的依据
compute maxphat = max(p0, p1, p2, p3, p4, p5).
compute predcls = 99.
if(maxphat = p0) predcls = 0.
if(maxphat = p1) predcls = 1.
if(maxphat = p2) predcls = 2.
if(maxphat = p3) predcls = 3.
if(maxphat = p4) predcls = 4.
if(rnaxphat = p5) predcls = 5.
freq /var = predcls.
CROSSTABS
    /TABLES = profread BY predcls
    /cells = count expected
    /FORMAT = AVALUE TABLES.
```

附录四 | 第 6 章

D1. 相邻类别(AC)模型

```
libname sagebook "C:\My Documents\research\ordinal new\sagebook\ordinal sas stuff";

proc freq data = sagebook.gonomiss;
    tables plageent profread* gender;
run;

proc catmod data = sagebook.gonomiss;
    direct gender;
    response alogit out = acgender;
    model profread = _response_ gender /wls pred;
run;

proc contents data = acgender;
run;

data go; set acgender;
    odds = exp( - pred_) ;
    predprob = odds /(1 + odds);
run;

proc freq data = go;
    tables gender* _number_* predprob;
run;
```

D2. 虚无模型

```
proc catmod data = sagebook.gonomiss;
    population gender;
    response alogit out = nogender;
    model profread = _response_ /wls pred;
run;
```

D3. 二变量模型

```
data g02; set sagebook.gonomiss;
     if plageentle 62 then agecat = 1;
     if plageent gt 62 AND plageentle 68 then agecat = 2;
     if plageent gt 68 AND plageentle 74 then agecat = 3;
     if plageent gt 74 then agecat = 4;
run;

proc catmod data = go2;
     direct gender agecat;
     response alogit out = acfull;
     model profread = _response_ gender agecat /wls pred;
run;

proc contents data = acfull;
run;

data go3; set acfull;
     odds = exp(_pred_);
     predprob = odds /(1 + odds);
     obsodds = exp(_obs_);
     obsprob = obsodds /(1 + obsodds);
run;

proc corr data = go3;
     var obsprob predprob;
     var _obs_ _pred_;
run;
```

注释

[1] 进一步了解测量层次问题，请参阅克利夫(Cliff, 1996b)以及阿格雷斯提和费雷(Agresti & Finlay, 1997)。

[2] 取得公共使用的 ECLS-K 数据的更多信息在网上可以查到：http://nces.ed.gov/ecls。

[3] 使用的是修改过的熟练度分数(C1RRPRF1 到 C1RRPRF5，等等)。

[4] ECLS-K 熟练度水平遵循一个古特曼模型(Guttman model)，也就是通过某一特定技能的儿童被认为是掌握了所有较低的技能水平。在一年级秋季和春季数据中，仅有 5.5%的儿童的阅读没有遵循这种模式，6.6%的儿童的数学没有遵循该模式。NCES(2002)报告认为，这种模式也许更多是意味着这些学生是猜测的，而非意味着一种不同的习得顺序。在 ECLS-K 数据的三年级里，由 NCES 所决定的各个儿童最高熟练度分数现在直接包括在数据集中；早些时候的数据集，包括本书中使用过的，仅包含了二分的熟练程度变量。对于响应模式不遵从古特曼模型的小部分学生，NCES 将他们的最高熟练度报告为“缺失”。然而，对于这里使用的数据，作者对某一儿童掌握(或者没有掌握)某一最高定序熟练度分数的指派是依据记录的 4 题答对 3 题的标准。

[5] 如一位评阅者提出的，将 ECLS-K 数据中得到的定序分数概念化为计数的过程也是可能的，也就是，一个计算儿童所通过的熟练度类别的个数(0 到 5，对于 ECLS-K 的一年级数据)。计数数据的模型，例如泊松或者负二项，也许是本书讨论的定序 logistic 模型的替代方法。然而，对于本书中分析的 ECLS-K 数据，这些策略也有一些局限，包括泊松过程中事件之间的独立性假设，不同学生间掌握度的同质性假设，一年级数据可能数目的上限等。随着三年级数据集的释出，将熟练度类别扩展到 13(包括原始的 0 到 5 级熟练度)，泊松回归过程的应用也许是理解影响儿童学习的因素的另外一种方法。如果能够参考下朗(Long, 1997)等人提到的对泊松回归过程的发展和应用信息中的不同的计数过程，那些基于定序数据研究的读者也许能够概念化得更好。

[6] 变异和饱和模型的讨论在这里多少有些简化。实际上，的确有若干定义饱和模型的方法，得到不同的变异值。

[7] 参见霍斯默和莱默苏(2000：147—156)以获得更多的关于 H-L 测试效能的信息，也有关于其他模型拟合检验的讨论。

[8] 作为关联的测量，Somers'D 既有对称形式也有非对称形式。SAS 的 LOGISTIC 过程计算并展示 $D_{x.y}$ 而非 $D_{y.x}$，也就是，计算 Somers'D 时预测概率被作为因变量对待，篇幅原因不再详述该统计量不同形式间的差异。细节可以参考利贝特劳(Liebetrau, 1983)、彭和尼科尔斯(Peng & Nichols, 2003)，以及彭和苏(Peng & So, 1998)等人的研究。

参考文献

AGRESTI, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*(*2*), 290—301.

AGRESTI, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

AGRESTI, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.

AGRESTI, A., & FINLAY, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

ALLISON, P. D. (1995). *Survival analysis using SAS: A practical guide*. Cary, NC: SAS Institute.

ALLISON, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.

ANANTH, C. V., & KLEINBAUM, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26* (*6*), 1323—1333.

ANDERSON, J. A. (1984). Regression and ordered categorical variables[with discussion]. *Journal of the Royal Statistical Society*, *Series B*, 46, 1—40.

ARMSTRONG, B. G., & SLOAN, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, 129(1), 191—204.

BENDER, R., & BENNER, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, 42(6), 677—699.

BENDER, R., & GROUVEN, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10), 809—816.

BOROOAH, V. K. (2002). *Logit and probit: Ordered and multinomial models*. Thousand Oaks, CA: Sage.

BRANT, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171—1178.

Center for the Improvement of Early Reading Achievement. (2001). *Put reading first: The research building blocks for teaching children to read, Kindergarten through grade 3*. Washington, DC: Government Printing Office.

CIZEK, G. 1., & FITZGERALD, S. M. (1999). An introduction to logistic regression. *Measurement and Evaluation in Counseling and Development*, 31, 223—245.

CLIFF, N. (1993). What is and isn't measurement. In G. Keren & C. Lewis(Eds.), *A handbook for data analysis in the social and behavioral sciences: Methodological issues* (pp. 59—93). Hillsdale, NJ: Lawrence Erlbaum Associates.

CLIFF, N. (1994). Predicting ordinal relations. *British Journal of Mathematical and Statistical Psychology*, 47, 127—150.

CLIFF, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multi-*

variate Behavioral Research, 3(3), 331—350.

CLIFF, N. (1996b). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

CLIFF, N., & KEATS, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

CLOGG, C. C., & SHIHADEH, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

COLE, S. R., & ANANTH, C. V. (2001). Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International Journal of Epidemiology*, *30*, 1379—1382.

COX, C. (1988). Multinomial regression models based on continuation ratios. *Statistics in Medicine*, 7, 435—441.

COX, D. R. (1972). Regression models and life tables[with discussion]. *Journal of the Royal Statistical Society B*, 74, 187—220.

DEMARIS, A. (1992). *Logit modeling* (Quantitative Applications in the Social Sciences, No. 86). Newbury Park, CA: Sage.

FOX, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

GIBBONS, J. D. (1993). *Nonparametric measures of association* (Quantitative Applications in the Social Sciences, No. 91). Newbury Park, CA: Sage.

GOLDSTEIN, H., RASBASH, J., PLEWIS, I., DRAPER, D., BROWNE, W., YANG, M., et al. (1998). *A user's guide to MLwiN*. London: Multilevel Models Project, Institute of Education, University of London.

GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537—552.

GOODMAN, L. A. (1983). The analysis of dependence in cross-classifications having ordered categories, using loglinear models for frequencies and log-linear models for odds. *Biometrics*, 39, 149—160.

GREENLAND, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13, 1665—1677.

GRISSOM, R. J. (1994). Statistical analysis of ordinal categorical status after therapy. *Journal of Consulting and Clinical Psychology*, 62(2), 281—284.

GUTTMAN, L. A. (1954). A new approach to factor analysis: The radix. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258—348). New York: Columbia University Press.

HALL, G. E., & HORD, S. M. (1984). *Change in schools: Facilitating the process*. Albany: State University of New York Press.

HEDEKER, D., & GIBBONS, R. D. (1996). MIXOR: A computer program for mixed effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 57—176.

HEDEKER, D., & MERMELSTEIN, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, *33*(4), 427—455.

HOSMER, D. W., & LEMESHOW, S. (1989). *Applied logistic regression*. New York: John Wiley & Sons.

HOSMER, D. W., & LEMESHOW, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons.

HUYNH, C. L. (2002, April). *Regression models oj ordinal response data: Analytic methods and goodness-of-fit tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

ISHII-KUNTZ, M. (1994). *Ordinal log-linear models* (Quantitative Applications in the Social Sciences, No. 97). Thousand Oaks, CA: Sage.

JENNINGS, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81, 471—476.

JOHNSON, R. A., & WICHERN, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

JORESKOG, K. G., & SORBOM, D. (1996). *LISREL* 8 *user's reference guide*. Chicago: Scientific Software International.

KNAPP, T. R. (1999). Focus on quantitative methods: The analysis of the data for two-way contingency tables. *Research in Nursing and Health*, 22, 263—268.

KOCH, G. G., AMARA, 1. A., & SINGER, J. M. (1985). A two-stage procedure for the analysis of ordinal categorical data. In P. K. Sen(Ed.), *Biostatistics: Statistics in biomedical, public health and environmental sciences* (pp. 357—387). Amsterdam: North Holland.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., & TVERSKY, A. (1971). *Foundations of measurement: Vol. I. Additive and polynomial representations*. New York: Academic Press.

LAARA, E., & MATTHEWS, J. N. S. (1985). The equivalence of two models for ordinal data. *Biometrika*, 72(1), 206—207.

LIANG, K. Y., & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13—22.

LIAO, T. F. (1994). *Interpreting probability models* (Quantitative Applications in the Social Sciences, No. 101). Thousand Oaks, CA: Sage.

LIEBETRAU, A. M. (1983). *Measures of association* (Quantitative Applications in the Social Sciences, No. 32). Beverly Hills, CA: Sage.

LONG, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

LONG, J. S., & FREESE, J. (2003). *Regression models for categorical dependent variables using STATA* (rev. ed.). College Station, TX: Stata.

McCULLAGH, P. (1980). Regression models with ordinal data[with discussion]. *Journal of the Royal Statistical Society*, *B*, 42, 109—142.

McCULLAGH, P., & NELDER, 1. A. (1983). *Generalized linear models*. London: Chapman and Hall.

McCULLAGH, P., & NELDER, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall/CRC Press.

McFADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105—142). New York: Academic Press.

MENARD, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

MENARD, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1), 17—24.

National Center for Education Statistics. (2000). *America's kindergarteners*. Retrieved from www.nces.ed.gov/pubsearchlpubsinfo.asp? pubid=2000070

National Center for Education Statistics. (2002). *User's manual for the ECLS-K first grade public-use data files and electronic codebook*. Retrieved from www.nces.ed.gov/pubsearch/pubsinfo.asp? pubid=2002135

NESS, M. E. (1995). Methods, plainly speaking: Ordinal positions and scale values of probability terms as estimated by three methods. *Measurement and Evaluation in Counseling and Development*, 28, 152—161.

O'CONNELL, A. A. (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, 33(3), 170—193.

O'CONNELL, A. A., McCOACH, D. B., LEVITT, H., & HORNER, S. (2003, April). *Modeling longitudinal ordinal response variables for educational data*. Paper presented at the 84th annual meeting of the American Educational Research Association, Chicago, IL.

PAMPEL, E. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

PENG, C. Y. J., & NICHOLS, R. N. (2003). Using multinomial logistic models to predict adolescent behavioral risk. *Journal of Modern Applied Statistical Methods*, 2(1), 1—13.

PENG, C. Y. J., & SO, T. S. H. (1998). If there is a will, there is a way: Getting around the defaults of PROC LOGISTIC. In *Proceedings of the MidWest SAS Users Group 1998 Conference* (pp. 243—252). Retrieved from http://php.indiana.edu/-tso/articles/mwsug98.pdf.

PETERSON, B. L., & HARRELL, E. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39(3), 205—217.

PLOTNIKOFF, R., BLANCHARD, C., HOTZ, S., & RHODES, R. (2001). Validation of the decisional balance scales in the exercise domain from the transtheoretical model. *Measurement in Physical Education and Exercise Science*, 5(4), 191—206.

PROCHASKA, J. O., & DICLEMENTE, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model. *Journal of Consulting and Clinical Psychology*, 51(3), 390—395.

PROCHASKA, J. O., & DICLEMENTE, C. C. (1986). Towards a comprehensive model of change. In W. R. Miller & N. Heather (Eds.), *Treating addictive behaviors: Processes of change* (pp. 3—27). New York: Plenum.

PROCHASKA, J. O., DICLEMENTE, C. C., & NORCROSS, J. C. (1992). In search of how people change: Applications to addictive behavior. *American Psychologist*, *47(9)*, 1102—1114.

RAUDENBUSH, S., BRYK, A., CHEONG, Y. E, & CONGDON, R. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

SIMONOFF, J. S. (1998). Logistic regression, categorical predictors, and goodness of fit: It de-

pends on who you ask. *American Statistician*, 52(1), 10—14.

SINGER, J. D., & WILLETT, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

SNOW, C. E., BURNS, M. S., & GRIFFIN, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

STARK, M. J., TESSELAAR, H. M., O'CONNELL, A. A., PERSON, B., GALAVOTTI, C., COHEN, A., et al. (1998). Psychosocial factors associated with the stages of change for condom use among women at risk for HIV/STDs: Implications for intervention development. *Journal of Consulting and Clinical Psychology*, 66(6), 967—978.

Statistical Analysis System. (1997). *SAS/STAT software: Changes and enhancements through release* 6.12. Cary, NC: Author.

Statistical Analysis System. (1999). *SAS Onlinedoc Version* 8. Retrieved from http://v8doc.sas.com/sashtml/

STEVENS, S. S. (1946). On the theory of scales of measurement. *Science*, *103(2684)*, 677—680.

STEVENS, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens(Ed.), *Handbook of experimental psychology*(pp. 1—49). New York: Wiley.

STOKES, M. E., DAVIS, C. S., & KOCH, G. G. (2000). *Categorical analysis using the SAS System*(2nd ed.). Cary, NC: SAS Institute.

TABACHNICK, B. G., & FIDELL, L. S. (2001). *Using multivariate statistics*(4th ed). Boston: Allyn & Bacon.

VAN DEN BERG, R., SLEEGERS, P., GEIJSEL, E, & VANDENBERGHE, R. (2000). Implementation of an innovation: Meeting the concerns of teachers. *Studies in Educational Evaluation*, 26, 331—350.

WEST, J., DENTON, K., & GERMINO-HAUSKEN, E. (2000). *America's kindergarteners: Findings from the early childhood longitudinal study, kindergarten class of* 1998—1999: *Fall* 1998(NCES 2000—070). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

ZILL, N., & WEST, 1. (2001). *Entering kindergarten: A portrait of American children when they begin school*(NCES 2001—035). Washington, DC: National Center for Education Statistics.

译名对照表

adjacent categories(AC) model	相邻类别模型
model fit assessment	模型拟合评估
assumption of parallelism	平行性假设
assumption of proportional or parallel odds	比例的或者平行的比数假设
Chi-square statistic	卡方统计量
choice of link and equivalence of two clog-log models	关联的选择和两种互补双对数模型的等价
clog-log link function	互补双对数关联函数
clog-log link model	互补双对数关联模型
conditional probabilities	条件概率
conditionally independent	条件独立
continuation ratio(CR)	连续比例
continuation ratio model	连续比例模型
cumulative odds(CO)	累积比数
Directionality of responses and formation of continuation ratios	响应的方向性和连续比例的形成
early childhood longitudinal study	儿童早期追踪研究
equal slopes assumption	等斜率假设
equal slopes assumption test	等斜率假设检验
full adjacent categories(AC) model	全相邻类别模型
full-model analyses	全模型分析
Full-model analysis for cumulative odds	累积比数的全模型分析
full-model continuation ratio analyses	全模型连续比例分析
generalized estimating equations(GEE)	一般估计方程
goodness of fit	拟合优度
graphical methods	做图方法
hazard(s)	风险
illness severity categories	疾病严重度类别
interval-level	定距水平
linearity and proportional odds	线性和比例比数
link functions	关联函数
logistic regression	logistic 回归
log-likelihood(LL)	对数似然
maximum likelihood(ML) estimates	极大似然估计
McFadden's pseudo R2	麦克法登虚拟 R2
model fit statistics	模型拟合统计量

multilevel structured	多层次结构化的
multinomial models	多分类模型
multiple regression(MR) analysis	多元回归分析
nominal-level	名义水平
null models	虚无模型
odds ratios	比数比
ordinal data	定序数据
parallel slopes assumption	平行斜率假设
partial proportional odds(PPO) model	偏比例比数模型
phonemic awareness	语音语韵觉识
probabilities of interest	兴趣概率
proficiency scores	熟练度分数
proportional hazards model	比例风险模型
ratio-level	定比水平
residual diagnostic	残差诊断
Somers' D	萨摩的 D 系数
stages-of-change models	阶段改变模型
absolute zero-points	绝对的零点

logit 与 probit：次序模型和多类别模型

作 者 简 介

瓦尼·布鲁雅(Vani Borooah)

英国阿尔斯特大学应用经济学教授。在 1987 年前,曾任剑桥大学应用经济学系高级研究员和剑桥大学王后学院经济学讲师。其研究关注于贫困、不平等和劳动力市场后果,并就这些议题发表了若干学术论文;尤其关注在经济和社会后果方面组间差异的公共政策含义。

译 者 简 介

张卓妮

中国人民大学社会学法学学士,2008 年香港科技大学社会科学部硕士,现为香港科技大学社会科学部博士候选人。研究方向为:社会分层与流动、中国的国内人口流动以及跨国流动。

序

要使普通最小二乘法(OLS)产生最优线性无偏估计(BLUE),必须符合经典回归假设。这些假设中有些假设比其他假设更容易实现。此外,违反这些假设的实际后果因假设的不同而不同。其中一个假设难以实现,那就是当对 OLS 的解释造成的严重后果没有解决时,假设因变量是连续的。相反,如果因变量是离散的,由两个或更多的结果类别构成,那么 OLS 就会产生严重的推论问题。在这种情况下,最大似然(maximum likelihood)技术比 logit 或 probit 通常更有效。

本书比较独特,因为它完全致力于分析当因变量具多类别时的估计情况。在概论之后,作者关注了具离散和次序形式的因变量。比如,假设某位政治科学家有选举调查的数据,并希望解释政治兴趣这一因变量,其中受访者的得分:0 = 低,1 = 中等,2 = 高。这个变量是离散的,受访者处于这三种类别中的一种。此外,这个变量是从“低”到“高”排序的。在这种有序变量情况下,我们可以说某个得分为“高”的人比某个得分为“低”的人具有更多的政治兴趣,但我们不能确切地说多多少。所以,OLS 回归看起来较不可取,而次序 logit 或次序 probit 更可取,因为它们适合这种较低的测量水平。布鲁雅(Borooah)教授详尽地阐释了这两种方法,试图解释社会剥夺(用三个类别测量,“没有被剥夺”、“轻度被剥夺”、“严重被剥夺”) 在不同个体间的差异。一个经常出现的问题是 logit 是否比 probit 更优,或者反之。这两种方法间根本上的理论差异涉及误差项的分布是逻辑分布还是正态分布。实际上,正如本书指出的,很难提供足够的理由说明为什么选择其中一种方法而非另外一种。

最终,本书把处理对象扩展到具两个以上结果的多类别或非次序因变量。比如,宗教的选择、住宅区的选择、购物中心的选择、工作的选择等。多类别 logit 的一个关键假设是无关选项独立性(IIA)的。正如布鲁雅教授所论述的,这个假设既是此技术的优点又是其缺点。他还对比数比(odds-ratios)和风险比(risk-ratios)做了重要但往往被忽视的区分。在二分类 logit 中,这两种比率之间没有差异,但是,在多类别 logit 中,结果是以风险比的方式显示的。

本书结尾给出了非常有用的实际用于产生书中表格结果的计算机程序详情。这种

逐步对计算机程序进行注释的方式让读者明白如何运行数据分析。讲解中具体使用的软件是 STATA，但作者还指出了 SAS，SPSS 和 LIMDEP 中其他可用的程序。总体而言，这本书为估计和解释从更复杂的离散因变量模型中得到的结果提供了一个可行的指南。

——迈克尔·刘易斯-贝克(Michael Lewis-Beck)

第 1 章 | 概论

戏剧评论家肯尼思·泰南(Kenneth Tynan)曾出色地描述过他的职业:由“知道路线,但是不知道怎么开车”的人组成。社会科学研究者经常发现自己处于类似的情形之中。有些人能在研究蓝图中定位,却不知如何从一个地方到另一个地方。其他人可以在研究中驱车前进却无法超越固定的路线。只有少数幸运的人既能导航又能飞行。对这种现象的察觉界定了本书的大概意图:(1)探寻穿越次序和多类别 logit 模型“地形”的主要路径;(2)为在这些路径中驱车前进提供指引。但首先,在发动机器之前,我们先对其简单介绍一下。

很多时候,适合我们分析的变量仅仅只是对定性结果的一种编码。这些模型就是定性选择模型。举例来说,在评判某种政府行为时,一个人非常赞成(编码 = 1),赞成(编码 = 2),反对(编码 = 3),或者强烈反对(编码 = 4);或者一个人投自由党的票(编码 = 1),投保守党的票(编码 = 2),或者工党的票(编码 = 3)。在这种情形之下,因变量取的是一组离散的、互斥的、完全穷尽的值。这与其他至少在理论上假设因变量的取值是连续数值的情况不同。虽然传统的回归方法不适合离散因变量的统计分析,但是我们可以用回归分析的原理,建立把观察到的结果与特定“决定性的”或“解释性的”变量连接起来的模型。因变量取值大于 2 个的定性选择模型为多元结果模型。我们还可以进一步把多元结果划分为:次序结果(如前述对政府行为的赞成程度的例子)和非次序结果(如前述投票的例子)。

定性选择模型在应用计量经济学分析中已经成为“成长型产业”。社会科学家一直对在互斥的可选项中进行选择的问题感兴趣。越来越多可利用的调

查数据(横截面或面板格式)意味着越来越多的学者可以把他们的知识思索转变成实实在在的成果。学者们带着一系列具体的问题从这些数据中寻找答案。然后,当分析完这些数据之后,结果往往隐藏在关于什么才是"正确"答案的模棱两可之中。为了驱散这种模糊不清的迷雾,必须对教科书技术进行巧妙的处理,以使推导出的结果明确地指向正确的方向。本章在社会经济不平等相关问题的背景之下,讨论了对多元结果模型——包括次序和非次序结果——的估计、模拟和解释。

次序和非次序模型对各自的分析有不同的技术要求。次序模型可以由logit 方法即次序 logit 模型,或者 probit 方法即次序 probit 模型进行估计。非次序结果模型最容易由 logit 方法估计,虽然原则上用 probit 的方法也可以对这种模型进行估计,但是由于计算方面的原因,这常常并不可行。因此,非次序多元结果模型被称为多类别 logit 模型。

多类别 logit 模型可以是有条件的,就是说在不同选项之间进行的选择不仅仅取决于做决定的个人所具有的特征,同时还取决于选择本身的属性。举个例子,个人选择惠顾哪个购物中心可能取决于购物中心本身的属性(商店的数量与种类,中心的保养标准),这一属性并不随个人情况而变;同时也取决于收入和家庭大小,这些随个人而变。一个更复杂的因素是个人特征和选择属性之间还可能互相影响:对购物中心的选择可能决定于个人从居住地(居住地属于个人的特征)到某个特定的购物中心(购物中心的位置属于选择的属性)的距离。次序模型第 2 章中论述,多类别模型将在第 3 章中论述,这两章具有两个共同的目的:(1)传达对这些模型基本方法论的理解;(2)传授在社会科学研究中运用这些模型的技能。

满足第一个目的需要在过度单纯化与过分技术性之间寻找一个安全的通道。作者已经尝试着在这条狭窄的通道中航行并避免船破水穿。然而,为了减轻它的负担,作者假设读者已经掌握了足够的理解本书的知识。其中,这些知识包括:

(1) 了解线性概率模型的不足之处,或者为什么普通的回归分析并不适用于具离散因变量的模型。

(2) 熟悉应用于因变量只有两个可能结果的模型:logit 和 probit 方法。

对于那些希望重温这些话题的读者来说,这个系列里奥尔德里奇和纳尔逊(Aldrich & Nelson, 1984)、德马里斯(DeMaris, 1992)、廖(Liao, 1994)和梅纳德(Menard, 1995)等人的专著,以及更具概括性的计量经济书如格林(Greene)2000 年的著作,提供了非常好的综述。

满足第二个目的在作者看来更加困难。对一个模型的运用需要好几个认识层次。首先,我们必须非常清楚要提的问题。其次,如何回答它们也是一个问题。通常,一个特定的研究问题可以用不止一种方法来回答,所以知道这些不同的方法怎么不同,为什么不同,以及眼前什么方法可能是解决问题的最好方法就很重要。最后,当我们已经决定了要提的问题以及怎么回答它们,就要处理获得答案的实践问题——通过取得并解释分析结果来执行研究策略。

作者已经尝试着解决这些议题,即在次序 logit 和 probit,以及多类别 logit 模型的框架中采取一个三重策略。第一,作者试图在对方法论与经验分析的阐释中实现尽可能多的交集。几乎每一章的理论总结都可在随后的应用内容中找到对应的分析。

第二,作者试图用两个固定的应用项目来进行阐释。这样做是因为作者觉得用一个单一的经验("真实世界")线索贯穿全章比用相互脱节的例子分开阐释可能更有利于说明一个模型的应用。

第 2 章次序 logit 和 probit 以对社会剥夺状态的经验分析为基础。这个项目使用的数据来自北爱尔兰对将近 14000 个人进行的关于社会剥夺的根源以及天主教徒与新教徒之间被剥夺经历的不平等状况的普查。第 3 章多类别 logit 模型基于对英国少数民族的职业获得的经验研究。这个项目使用的数据来自英国对将近 10 万全职男性雇员进行的普查,旨在研究黑加勒比人、印第安人和白人处于各种职业类别时的不同机会。

第三,最后一章列举了用于产生经验结果的完整的电脑程序清单。据作者所知,至少有 4 种众所周知的具高度好评的软件,这些软件除了能解决其他方面的问题,还很适用于解决与本书的讨论内容类似的问题:SAS, SPSS v10.0(印第安纳大学统计与数学计算中心为在次序与非次序结果事件的分析中应用 SAS

和 SPSS 提供了很好的入门课程[1]),LIMDEP(Greene, 1995)以及 STATA (STATA, 1999)。由于作者对 STATA 比较熟悉,因此第 4 章中的程序均用 STATA 编写。几乎每一行命令都附有一个注解,以说明这些命令应起的作用以及其如何与前面章节中的材料相关。

第 2 章 | 次序模型

简介

假设有 N 个人(标号为 $i=1, \cdots, N$),每个人都会发生一个"事件"。假设这个事件有 $M>2$ 种结果,标号为 $j=1, \cdots, M$,这些结果是相互排斥且完全穷尽的。令变量 Y_i 取的值代表第 i 个人的结果:$Y_i=1$ 表示这个人出现了第一种结果 ($j=1$),$Y_i=2$ 表示这个人出现了第二种结果 ($j=2$),如此这般,直至 $Y_i=M$ 表示出现最后一种结果 ($j=M$)。进一步假设这些结果在本质上就是次序的,意味着与变量 Y_i 更高的取值相关的结果比与其更低的取值相关的结果排序更高。或者说,因变量 Y_i 与有序的结果相关:"更强的"结果与因变量更高的值联系在一起。但是,这些结果序号间的差异并不表示它们在程度上的差异;与 $Y_i=2$ 相关的结果并不是与 $Y_i=1$ 相关的结果的两倍强度。因此,一个有序因变量的实际取值之间是不相关的,它只意味着更大的取值对应于更强的结果:我们可以令 $Y_i=5$ 表示发生了第一种结果,令 $Y_i=7$ 表示发生了第二种结果,以此类推。

一个人的健康状况可以提供一个次序结果的例子。"差"、"良好" 以及"非常好"的健康状况可以用一个取值分别为 1, 2 和 3 的变量表示。在这个例子中,$Y_i=3$(健康状况非常好)表示其结果比 $Y_i=2$(健康状况良好)更好,依次,后者又比 $Y_i=1$(健康状况差)表示的结果更好。其他次序结果的例子,如个人的保险范围(没有、半保、全保),或者处于工作年龄的个人之雇佣状况(不从事经济活动、失业、雇员)。而个人的宗教信仰则可提供非次序变量的例子,如 $Y_i=1$ 表示新教徒,$Y_i=2$ 表示犹太教徒,$Y_i=3$ 表示穆斯林,$Y_i=4$ 表示印

第安教徒。虽然此例中的这些结果都不一样，但是它们却不可以排序，因此不能被视为有序结果。换一种方式说，这个表示宗教信仰的因变量是非次序的。

当结果明显具有次序的时候，我们可以考虑这样一种实际情况：因变量既是离散的又是有序的。例如，若结果被编码为 1，2，3，线性回归就会把 3 与 2 之间的差异与 2 与 1 之间的差异同等对待，尽管这些数字其实只是一种排列方式而并没有什么重要意义。另一方面，用多类别 logit 的方法对一个有序因变量进行计量经济学关系的估计，则意味着这些数据的次序性所传达的信息被舍弃掉了。用于估计因变量多于两个结果[2]且这些结果同时是离散的又是有序的模型，最常用且适当的方法就是次序 logit 和次序 probit 的方法[3]。

但是，以上所述受限于一个重要的附加说明。次序 logit 和 probit 的一个关键假设是平行斜率假设。这个假设的含义将在下面进行略为详细的讨论，但大体上，它的意思是如果有一个变量影响了一个人处于这些次序类别中的哪一个结果的可能性（如饮食对健康状况的影响），就假定这个变量与结果间相关联的系数对所有的结果是一样的（特定的饮食对一个人处于非常好的健康状况的可能性的影响与对他或她处于差的健康状况的可能性的影响是一模一样的）。如果这个假设不成立，与某一特定变量相关联的斜率系数对不同的结果是不一样的（特定的饮食对一个人处于非常好的健康状况的可能性的影响不同于对他或她处于差的健康状况的可能性的影响），那么次序 logit 和 probit 的方法就不再适当，这种模型就应该用多类别 logit 的方法。

我们并不是总能明确断定结果是有序的（还是无序的），这样一种事实给出了另外一个谨慎使用次序估计方法的原因。例如，一个人居住于城市的哪个位置（北边，南边，东边或者西边）表面上看是一个非有序的变量。但是如果我们知道城市的某些位置相比其他位置具有更健康的居住条件，那么界定个人居住位置的变量就需要有序的内涵，比如，住在北边（$Y_i = 4$）比住在南边（$Y_i = 3$）要好。但是，当并不确定某个变量是次序的还是非次序的时候，一个明智的准则是把它当成非次序的来对待，并用多类别 logit 的方法对以它为因变量的模型进行估计。这个准则是切合实际的，因为如果把一个事实上是非次序的结果变量当成次序的来对待，就给这些其实无序的结果强加了一个排序并提出了平行斜

率(参考上文,并于下文有详细讨论)的限制性假设,而这样很可能会使估计发生偏差。另一方面,如果把次序结果变量当成非次序的,未能对这些结果给予合理的排序,会损失估计的效率,但却不可能使估计发生偏差。在这两种可能出现的错误之中,效率的损失与有偏的估计相比并不那么严重。

这种不确定性的另一个例子,同时也是下一章多类别 logit 中的应用实例,是个人的职业。一个人从事的职业是非技术的、半技术的还是技术的;以及其工作是专业类别还是管理类别,都可以看做是一种个人选择,尽管影响这个选择的限制条件可能因人而异且尤其随种族和/或性别而改变。在解释这种选择的时候,职业结果可以看做是"非次序的",意即一种职业类别相比另一种职业类别并非固有地更理想或更不理想。对于这种解释,分析职业结果的适当的估计方法是多类别 logit。其实,这正是施米特和斯特劳斯(Schmid and Strauss)(1975)对 1000 个人进行关于教育、经验、种族和性别的职业分析时所采用的方法。这种方法后来在格林(Greene)(2000:859)讨论多类别 logit 时作为例子被应用。

从另一方面来说,假如一个大学教授被问及他/她更喜欢一个银行家还是一个看门人作为他/她的女婿,那个教授更可能会选择前者。在表达这种偏好时,这个教授暗中对职业进行了排序,把银行家排在了比看门人更高的层次(选择了更理想的女婿)。但是,重点是这种排序纯粹是主观的[4](即银行家与看门人相比并不是固有地更理想),那个教授对于他/她的女婿应该更健康而非不健康的偏好并不具备客观性。这个故事的寓意在于,我们最好把结果当成是非次序的来对待,除非有好的理由对这些结果强加一种排序。[5]

方法论

用一个具体的例子来说明次序 logit 和 probit 模型的方法论及其根本逻辑可能是最好的办法。假设有 N 个人(标号为 $i = 1, \cdots, N$),住在一个地区,且每个人的"社会剥夺程度"(degree of deprivation)可以用变量 D_i 的取值来表示,且 D_i 越高的取值代表越高的剥夺程度。对某个特定个人赋予的由这个"剥夺指数"(deprivation index)假定的值——此后称做他/她的"剥夺分数"——取

决于这个人既有的多种因素。

这些因素举例来说可能包括失业、单亲家长及居住于某一特定区域。假设剥夺指数 D_i，是 K 个因素(决定变量)的一个线性函数，这 K 个因素的取值对于个人 i 来说，为 X_{ik}，$k = 1, \cdots, K$。这意味着剥夺指数可以表示为：

$$D_i = \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i = Z_i + \varepsilon_i \qquad [2.1]$$

其中，β_k 是与第 k 个变量($k = 1, \cdots, K$)相关的系数，$Z_i = \sum_{k=1}^{K} \beta_k X_{ik}$。如果 $\beta_k > 0$，对某个特定的个人第 k 个因素取值的增加会导致他/她剥夺分数的上升，反之 $\beta_k < 0$ 会导致分数的下降。但是，因为剥夺分数与导致剥夺的因素之间的关系并不精确——例如，可能有些因素并不包括在方程内或者有些因素的测量并不准确——所以此方程还包括了一个误差项 ε_i，用以捕捉这种不精确性。

方程 2.1 存在的问题是，用 D_i 的取值所表示的个人剥夺状态的精确图案是很难观察到的。剥夺指数是一个潜在变量，(虽然在概念上有用)但在原则上和实践上都难以察觉，方程 2.1 是一个潜在回归，按照现在的情况来看是无法进行估计的。

但是，我们能观察到的是一个人被剥夺的程度(deprivation level)——例如一个人可以被归类为“没有被剥夺”、“轻度被剥夺”或“严重被剥夺”——而且变量 Y_i 可以以下列方式与这些剥夺程度相关联，$Y_i = 1$ 如果此人没有被剥夺，$Y_i = 2$ 如果此人轻度被剥夺，$Y_i = 3$ 如果此人严重被剥夺。按照前面的讨论，Y_i 是一个有序变量。样本中个人在这三个剥夺程度中的分类内在地(implicitly)基于潜在变量 D_i 的值以及“临界”值 δ_1 和 δ_2，如下所示：

$$\begin{aligned} Y_i &= 1, \text{如果 } D_i \leqslant \delta_1 \\ Y_i &= 2, \text{如果 } \delta_1 \leqslant D_i \leqslant \delta_2 \\ Y_i &= 3, \text{如果 } D_i \geqslant \delta_2 \end{aligned} \qquad [2.2]$$

方程 2.2 中的 δ_1，$\delta_2 \geqslant 0$，是跟方程 2.1 中的 β_k 一起有待估计的未知参数($\delta_1 < \delta_2$)。一个人在剥夺程度上的归类取决于他/她的剥夺分数 D_i 是否跨过

某个临界值。Y_i 取值为 1，2 和 3 的概率表示如下：

$$\begin{aligned}\Pr(Y_i=1)&=\Pr(Z_i+\varepsilon_i\leqslant\delta_1)=\Pr(\varepsilon_i\leqslant\delta_1-Z_i)\\ \Pr(Y_i=2)&=\Pr(\delta_1\leqslant Z_i+\varepsilon_i\leqslant\delta_2)=\Pr(\delta_1-Z_i<\varepsilon_i\leqslant\delta_2-Z_i)\\ \Pr(Y_i=3)&=\Pr(Z_i+\varepsilon_i\geqslant\delta_2)=\Pr(\varepsilon_i\geqslant\delta_2-Z_i)\end{aligned} \qquad [2.3]$$

N 个观察中的每一次都被视为多项分布中的一次单一的抽取，在这种情况下，这个多项分布有三种结果，没有被剥夺、轻度被剥夺或严重被剥夺。假设 N 个人，N_1 没有被剥夺，N_2 轻度被剥夺，N_3 严重被剥夺。[6]那么观察到这个样本的可能性，就是个体观察值概率的乘积，为：

$$\begin{aligned}L&=[\Pr(Y_i=1)]^{N_1}[\Pr(Y_i=2)]^{N_2}[\Pr(Y_i=3)]^{N_3}\\&=[F(\delta_1-Z_i)]^{N_1}[F(\delta_2-Z_i)-F(\delta_1-Z_i)]^{N_2}\\&\quad\times[1-F(\delta_1-Z_i)]^{N_3}\end{aligned} \qquad [2.4]$$

其中，$F(x)=\Pr(\varepsilon_i<x)$ 是误差项的累积概率分布。如果我们已知误差项的概率分布——即我们知道什么是 $F(x)$——那么我们就可以选择像 β_k，δ_1，和 δ_2 的估计值一样的那些使观察到样本结果的可能性最大化的值[7]。若缺乏这样的信息，则我们可以假定误差项服从一个特定的概率分布。

次序 logit 与次序 probit 模型之间的差别在于：方程 2.1 中的误差项 ε_i 的(假定的)分布不同。次序 logit 模型是假定 ε_i 具逻辑分布的结果，而次序 probit 模型则为假定 ε_i 具正态分布的结果。很自然地，我们会问哪种分布更适合。[8]逻辑分布除了在末尾部分比正态分布大很多之外，其他与正态分布相似。[9]正如格林(2000:815)指出的："要在理论上证明选择一种而非另一种分布更为合理是很困难的事……在大部分应用中，(它们)看起来并没有什么大的差别。"

使用系数 β_k 的估计值 $\hat{\beta}_k$ 可以让我们对样本中的每个个体计算估计值 $\hat{Z}_i=\sum_{k=1}^{k}\hat{\beta}_kX_{ik}$。使用 $\hat{Z}_i$ 以及临界参数 δ_1，δ_2 的估计值 $\hat{\delta}_1$ 和 $\hat{\delta}_2$，可以对样本中每个个体处于不同剥夺程度的概率进行估计。这些估计可分别用 $\hat{P}_{i1}$，$\hat{P}_{i2}$ 和 $\hat{P}_{i3}$ 表示，计算如下：

$$\hat{P}_{i1}=\Pr(\varepsilon_i\leqslant\hat{\delta}_1-\hat{Z}_i)=F(\hat{\delta}_1-\hat{Z}_i) \qquad [2.5a]$$

$$\hat{P}_{i2}=\Pr(\hat{\delta}_1-\hat{Z}_i<\varepsilon_i\leqslant\hat{\delta}_2-\hat{Z}_i)=F(\hat{\delta}_2-\hat{Z}_i)-F(\hat{\delta}_1-\hat{Z}_i) \qquad [2.5b]$$

$$\hat{P}_{i3}=\Pr(\varepsilon_i\geqslant\hat{\delta}_2-\hat{Z}_i)=1-F(\hat{\delta}_2-\hat{Z}_i) \qquad [2.5c]$$

其中，$\sum_{j=1}^{3}\hat{p}_{ij}=1$ 对每个 $i=1,\cdots,N$。

上面描述的模型同时也称做比例比数模型(proportional-odds model)，因为如果我们考虑某个类别 $j=m$ 的比数比(OR)，$OR(m)=\frac{\Pr(Y_i\leqslant m)}{\Pr(Y_i>m)}$，那么这个比数比就独立于类别 m。比数比被假定为对所有的类别保持不变[10]。

关于符号的说明

用 STATA 估计的次序回归(不管是 logit 还是 probit)并不明显包含截距项:或者说,方程 2.1 中的 $\beta_k(k=1,\cdots,K)$全部都是斜率。截距项没有明显地表示出来,因为如下所示它被并入到临界值 δ_1 和 δ_2 中了。另一方面,格林(2000:876)在用公式表示次序回归时,明显地包括了截距项。对应于方程2.1，他的方程是:

$$D_i=\beta_0+\sum_{k=1}^{K}\beta_k X_{ik}+\varepsilon_i=\beta_0+Z_i+\varepsilon_i=W_i+\varepsilon_i \qquad [2.6]$$

其中 β_0是截距项,$W_i=\beta_0+Z_i$。格林的临界值用 μ_1 和 μ_2 表示(跟 STATA 里面的临界值 δ_1 和 δ_2 不一样,究竟如何不一样表现如下)。格林(2000)设第一个临界值 μ_1 等于 0。[11] 所以在他的表述中,表示概率 $\hat{P}_{i1}$，$\hat{P}_{i2}$和 $\hat{P}_{i3}$的方程为:

$$\hat{P}_{i1}=\Pr(\varepsilon_i\leqslant-\hat{W}_i)=F(-\hat{W}_i)=F(-\hat{\beta}_0-\hat{Z}_i)$$
$$=F(\hat{\delta}_1-\hat{Z}_i) \qquad [2.7a]$$

$$\hat{P}_{i2}=\Pr(-\hat{W}_i<\varepsilon_i\leqslant\hat{\mu}_2-\hat{W}_i)=F(\hat{\mu}_2-\hat{W}_i)-F(-\hat{W}_i)$$
$$=F(\hat{\mu}_2-\hat{\beta}_0-\hat{Z}_i)-F(-\hat{\beta}_0-\hat{Z}_i)$$
$$=F(\hat{\delta}_2-\hat{Z}_i)-F(\hat{\delta}_1-\hat{Z}_i) \qquad [2.7b]$$

$$\hat{P}_{i3} = \Pr(\varepsilon_i \geqslant \hat{\mu}_2 - \hat{W}_i) = 1 - F(\hat{\mu}_2 - \hat{W}_i)$$
$$= 1 - F(\hat{\mu}_2 - \hat{\beta}_0 - \hat{Z}_i) = 1 - F(\hat{\delta}_2 - \hat{Z}_i) \quad [2.7c]$$

当 $\delta_i = \mu_i - \beta_0$ 时，STATA 方程 2.5a 至 2.5c 与格林(2000)的方程 2.7a 至 2.7c 是相同的。[12]也就是说，STATA 的临界值等于格林的临界值减去截距项。从这种意义上说，STATA 把截距项合并到临界值中去了。[13]

次序 logit

在逻辑分布的情况下，随机变量 X 的累积分布函数是：

$$\Pr(X \leqslant x) = \Lambda(x) = \exp(x)/[1 + \exp(x)] = 1/(1 + \exp(-x)) \quad [2.8]$$

如果假设误差项服从逻辑分布，

$$\Pr(Y_i = 1) = \Lambda(\delta_1 - Z_i) = 1/[1 + \exp(Z_i - \delta_1)] \quad [2.9a]$$

$$\Pr(Y_i = 2) = \Lambda(\delta_2 - Z_i) - \Lambda(\delta_1 - Z_i)$$
$$= 1/[1 + \exp(Z_i - \delta_2)] - 1/[1 + \exp(Z_i - \delta_1)] \quad [2.9b]$$

$$\Pr(Y_i = 3) = 1 - \Lambda(\delta_2 - Z_i) = 1 - 1/[1 + \exp(Z_i - \delta_2)] \quad [2.9c]$$

则对 β_k，δ_1 和 δ_2 的估计通过使用逻辑分布函数 $\Lambda(.)$ 代替 $F(.)$ 最大化似然函数(方程 2.4)获得。

次序 probit

标准正态变量[14](SNV)X 的累积分布是：

$$\Pr(X < x) = \Phi(x) = \int_0^x (1/2\pi)\exp(-X^2/2)dX \quad [2.10]$$

如果假设误差项是 SNVs，

$$\Pr(Y_i = 1) = \Phi(\delta_1 - Z_i) \quad [2.11a]$$

$$\Pr(Y_i = 2) = \Phi(\delta_2 - Z_i) - \Phi(\delta_1 - Z_i) \quad [2.11b]$$

$$\Pr(Y_i = 3) = 1 - \Phi(\delta_2 - Z_i) \quad [2.11c]$$

则对 β_k，δ_1 和 δ_2 的估计通过使用正态分布函数 $\Phi(.)$ 代替 $F(.)$ 最大化似

然函数(方程 2.4)获得。

边际效应:连续变量

一个意料之中的问题是当其中一个影响结果的变量的值改变的时候,出现不同结果的概率是怎么变化的。例如,如果年纪是影响剥夺的一个因素,那么个人处于不同的剥夺程度(没有被剥夺、轻度被剥夺或严重被剥夺)的概率是怎么受他/她是大一岁还是小一岁影响的? 在逻辑分布下,X_{ik}(第 i 个人的第 k 个决定变量的值)的一个小变化对个人 i 的这三种概率的边际效应是:

$$\frac{\partial \Pr(Y_i=1)}{\partial X_{ik}} = \frac{d}{dZ_i}[\Lambda(\delta_1 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}} = -\Lambda'(\delta_1 - Z_i)\beta_k \quad [2.12a]$$

$$\frac{\partial \Pr(Y_i=2)}{\partial X_{ik}} = \frac{d}{dZ_i}[\Lambda(\delta_2 - Z_i) - \Lambda(\delta_1 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}}$$

$$= [\Lambda'(\delta_2 - Z_i) - \Lambda'(\delta_1 - Z_i)]\beta_k \quad [2.12b]$$

$$\frac{\partial \Pr(Y_i=3)}{\partial X_{ik}} = \frac{d}{dZ_i}[1 - \Lambda(\delta_2 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}} = \Lambda'(\delta_2 - Z_i)\beta_k \quad [2.12c]$$

在正态分布下,则为:

$$\frac{\partial \Pr(Y_i=1)}{\partial X_{ik}} = \frac{d}{dZ_i}[\Phi(\delta_1 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}} = -\Phi'(\delta_1 - Z_i)\beta_k \quad [2.13a]$$

$$\frac{\partial \Pr(Y_i=2)}{\partial X_{ik}} = \frac{d}{dZ_i}[\Phi(\delta_2 - Z_i) - \Phi(\delta_1 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}}$$

$$= [\Phi'(\delta_2 - Z_i) - \Phi'(\delta_1 - Z_i)]\beta_k \quad [2.13b]$$

$$\frac{\partial \Pr(Y_i=3)}{\partial X_{ik}} = \frac{d}{dZ_i}[1 - \Phi(\delta_2 - Z_i)]\frac{\partial Z_i}{\partial X_{ik}} = \Phi'(\delta_2 - Z_i)\beta_k \quad [2.13c]$$

其中 $\Lambda'(x) = d\Lambda(x)/dx$ 和 $\Phi'(x) = d\Phi(x)/dx$ 分别是逻辑分布和正态分布的概率密度函数。边际效应可以通过估计在相关点上的合适的密度函数并乘以相关的系数获得。举例,由方程 2.8,逻辑分布的密度函数是:

$$\Lambda'(x) = \frac{d}{dx}\left[\frac{\exp(x)}{1+\exp(x)}\right] = \frac{[1+\exp(x)]\exp(x) - [\exp(x)]^2}{[1+\exp(x)]^2}$$

$$= \Lambda(x)[1 - \Lambda(x)]$$

方程 2.12a 的边际效应是：

$$\Lambda'(\delta_1 - Z_i)\beta_k = \Lambda(\delta_1 - Z_i)[1 - \Lambda(\delta_1 - Z_i)]\beta_k$$
$$= \frac{1}{1+\exp(Z_i - \delta_1)}\left[1 - \frac{1}{1+\exp(Z_i - \delta_1)}\right]\beta_k$$

现在如果第 k 个决定变量的值增加一点且 $\beta_k > 0$，那么在 logit 和 probit 模型下，没有被剥夺的概率都一定会降低，因为通过方程 2.12a 和 2.13a，$\Pr(Y_i = 1)$ 的推导与 β_k 相反方向。同时在两个模型中，严重被剥夺的概率都会上升，因为通过 2.12c 和 2.13c，$\Pr(Y_i = 3)$的推导与 β_k 方向相同。但是我们并不清楚中间那个概率会发生什么变化。取决于其他两个概率怎么变化，轻度被剥夺的概率，$\Pr(Y_i = 2)$，可能上升、下降或者保持不变。[15]因此，给定某个决定变量的变化，不可能根据相关系数的方向对所有概率的变化方向做出推断。只有处于两个极端情形的概率的变化方向能被确定。由于这个原因，格林(2000:878)警告说"我们在解释这个模型的系数时必须非常小心，因为这是模型中最不明显的"。

边际效应：虚拟变量

然而，上述估计边际效应的方法只在决定变量为连续变量而非虚拟变量时适用。虚拟变量的作用应该通过比较虚拟变量取其中一个值时结果出现的概率和它取另外一个值的结果出现的概率进行分析，且这种比较要在其他变量的值保持不变的情况下进行。例如，假设雇佣状态是影响剥夺的一个因素，$X_{ik} = 1$ 表示一个人失业，$X_{ik} = 0$ 则表示他/她受雇。为了分析如果他/她从受雇转为失业如何影响其处于三种不同的剥夺程度的概率，首先使用方程 2.9a 至 2.9c，假设 $X_{ik} = 1$ 对 Z_i 进行估计(称为 Z_i^1)；如果应用 probit 模型，则使用方程2.11a 至 2.11c，计算 $Y_i = 1$(没有被剥夺)，$Y_i = 2$(轻度被剥夺)和 $Y_i = 3$(严重被剥夺)这三种概率。然后，保持其他决定变量的值不变，假设 $X_{ik} = 0$ 对 Z_i 进行估计(称为 Z_i^0)，重新计算以上三种概率。注意，由方程 2.1，$Z_i^1 = Z_i^0 + \beta_k$。这两组概率之间的区别就是当个人从受雇状态($X_{ik} = 0$)转为失业状态($X_{ik} = 1$)，或从失业转为受雇，对他/她处于不同剥夺程度的概率的作用。

平行斜率假设

次序 logit 和 probit 模型的一个关键假设是方程 2.1 的斜率系数 β_k 并不随着所关注的剥夺结果的变化而改变。也就是说，次序 logit 和 probit 模型符合一种平行斜率累积模型，这种模型在 logit 的情况下，采取下列形式：

$$\log\frac{p_1}{1-p_1}=\alpha_1+\sum_{k=1}^{K}\beta_k X_{ik}$$

$$\log\frac{p_1+p_2}{1-p_1-p_2}=\alpha_2+\sum_{k=1}^{K}\beta_k X_{ik}$$

$$\cdots\cdots$$

$$\log\frac{p_1+p_2+\cdots+p_{M-1}}{1-p_1-p_2-\cdots-p_{M-1}}=\alpha_M+\sum_{k=1}^{K}\beta_k X_{ik}$$

$$p_1+p_2+\cdots+p_M=1$$

其中，$p_j=\Pr(Y_i=j)$。通过对数据进行多类别 logit 模型估计，我们可以对平行斜率假设的有效性进行检验(见第 3 章)。多类别 logit 模型允许斜率系数 β_k 在不同的结果 $j=1, \cdots, M$ 之间不同。当次序 logit 模型估计 K 个系数时，多类型 logit 模型估计 $K(M-1)$ 个系数。[16] 如果 L_1 是从次序 logit 模型得出的概率值，L_2 是从多类别 logit 模型得出的概率值，那么我们就能计算 $2(L_2-L_1)$，并与 $\chi^2(K(M-1))$ 进行比较。注意，这并不是严格的似然比检验，因为次序 logit 模型并不嵌套在多类别 logit 模型之中。因此，这个检验只是提示性的：一个"非常大的"χ^2 值会是引起担心的依据，一个"中等大的"的则不是(STATA, 1990:480)。但是，如果我们确实有理由相信平行斜率假设并不有效，那么此模型就该用多类别 logit 的方法进行估计，就算事实上因变量是明显有序的。

应用：剥夺状态

界定与构造剥夺指数

有 N 个人，标号为 $i=1, \cdots, N$。一种条件被定义为"剥夺导致条件"

(DIC),如果这种条件的出现导致一个人经历剥夺。假设有 K 种 DIC,标号为 $k=1, \cdots, K$, I_{ik} 是关于 DIC k 及个人 i 的一个类别变量,$I_{ik}=1$ 表示 DIC 对第 i 个人出现了, $I_{ik}=0$ 表示没有出现。那么对个人 i 的剥夺程度,记为 D_i^* ,定义为:

$$D_i^* = \sum_{k=1}^{K} \alpha_k^* I_{ik} \qquad [2.14]$$

其中 $\alpha_k^* > 0$ 是赋予第 k 个 DIC 的权重,且独立于所考虑的个人。如果与个人 DIC 相关的权重定义为 $\alpha_k^* = 1 - p_k$,其中 p_k 代表条件 k 发生的频数,那么 α_k^* 包含了"相对剥夺"的概念,即一个特定的 DIC 被经历的频数越小,那么当它被经历的时候赋予它的权重就越大。这些权重的使用响应了德赛和沙(Desai & Shah, 1988)的工作,他们在重新检验汤森(Townsend, 1979)的原始数据时,从根本上主张被剥夺了几乎所有人拥有的东西要比被剥夺了很少人拥有的东西影响更大。

只在偶然的情况下,权重 α_k^* 会总和为 1(sum to unity)。但是,它们却可通过定义 $\alpha_k = \alpha_k^* / \Omega$ 标准化,其中 $\Omega = \sum_{k=1}^{K} \alpha_k^*$ 。在这种标准化之下,个人 i 的剥夺程度可以定义为:

$$D_i = D_i^* \Omega^{-1} = \left(\sum_{k=1}^{K} \alpha_k^* I_{ik}\right) \Omega^{-1} = \sum_{k=1}^{K} \alpha_k I_{ik} \qquad [2.15]$$

因为 D_i 只是 D_i^* 的一种简单的标量转换,所以可以通过使用 D_i,与使用 D_i^* 一样,获得个人在他们的剥夺程度上的相同排序。但是,在标准化权重方面,剥夺指数 D_i 比 D_i^* 提供了更好的解释,所以随后的分析将会采用 D_i 这种标准。因为,在定义上 $\sum_{k=1}^{K} \alpha_k = 1$,所以方程 2.15 暗示着 $0 \leqslant D_i \leqslant 1$:当 $D_i = 0$ 时,表示没有任何一种 DIC 出现,也就是当对所有的 $k = 1, \cdots, K$, $I_{ik} = 0$ 时;当 $D_i = 1$ 时,表示所有的 DIC 都出现,也就是当对所有的 $k = 1, \cdots, K$, $I_{ik} = 1$ 时。

构造剥夺指数的一个主要问题在于决定应该加入哪些 DIC。回顾关于构造剥夺指数的文献,诺兰和惠兰(Nolan & Whelan, 1996)指出,鉴赏力的作用是

一个主要的问题。如果在生活模式中观察到的差异可以大部分归因于偏好而非资源，那么在构造剥夺指数时，若某些条目不被包含在内，也不能说明是因为需求的原因。例如，皮亚乔德(Piachaud, 1987)强调了处于相似收入水平的家庭在剥夺分数间存在的巨大差异。即使我们可以把偏好从需求中分离出来，在选择把哪些条目放入到剥夺指数中时，研究者的重要性又是另一个更深层次的问题。这样，问题就在于剥夺的测量是应该单单基于个人自己的环境，还是应根据他/她的更宽泛的社会以及地理环境。例如，布鲁雅和卡克(Borooah & Carcach, 1997)注意到居住社区质量在决定人们对犯罪的害怕程度上的重要性。最后，如果我们可以取得一系列令人满意的指标，还有一个问题在于如何在构造综合指数时对这些指标进行加权。数据本身的限制在这些问题上起中心作用；我们只能从现成可用的数据中构建剥夺指数，而不能从我们希望可得到的数据中构建该指数。作者所构建的剥夺指数基于1991年人口普查①的数据，该数据包含了在此地居住的13164个人的生活环境的信息。[17] 比如，这些数据告诉我们在普查时个人是否：

住在一个没有任何成员通常有机会使用一辆小车或小巴的家庭中；

住在一个没有任何成员是挣钱者的家庭中；

住在一个家庭成员没有专用的室内厕所的家庭中；

住在一个没有中央暖气系统的房子里；

住在一个没有通到室内的公共供水的房子里；

住在一个没有与公共排水管道连通的房子里；

住在一个代表非永久住处的房子里；

住在一个居住人数与房间数量的比例大于1的房子里；

长年患有疾病，面临健康问题或者残疾，这些问题限制了他/她的日常活动或者可能做的工作。

① 北爱尔兰1991年的人口普查。——译者注

利用这些信息，对退休者和非退休者分别构建了剥夺指数。这两组指数都围绕着相同的一组 DIC 构建，但是对这些 DIC 赋予的权重随着个人是否退休而不同。正如上面提到的，与某个 DIC 相关的权重反映了相关 DIC 被经历的频数——频数越低，权重越大。但是，其中一些 DIC 被经历的频数在已退休样本和非退休样本之间存在很大差异，在对退休者和非退休者构造剥夺指数时使用不同的权重反映了这种差异。取决于对他/她计算出来的剥夺指数的值，每个人都被指定了三个剥夺程度——没有被剥夺、轻度被剥夺、严重被剥夺——中的一种，[18]与每个结果($j=1,2,3$)和每个人($i=1,\cdots,N$)相关的是有序因变量 Y_i 的如下取值：

$Y_i = 1$ 如果这个人没有被剥夺

$Y_i = 2$ 如果这个人轻度被剥夺

$Y_i = 3$ 如果这个人严重被剥夺

应该特别强调的是，正是这些关于 Y_i 的值的数据典型地可供研究者分析利用。在 13164 个人的全部样本中，45.9%(6042 个人)被归为没有被剥夺，34.9%(4594 个人)被归为轻度被剥夺，19.2%(2528)被归为严重被剥夺。

方程设定

用于"解释"个人的剥夺程度的决定变量为：

AGE_i，以年为单位，通过设定 16 岁的人 $AGE_i = 0$ 进行标准化；[19]

$HIGHED_i = 1$，如果这个人具有英国标准的学士或者更高的学位资格；否则 $HIGHED_i = 0$；

$MIDED_i = 1$，如果这个人有高于 A 程度文凭(post-A level)①，但低于学位文凭；[20]否则 $MIDED_i = 0$；

① 英国文凭高级水平(A-level)是中学和大学的衔接课程，其考试成绩是从中学升入大学的考核标准；在这个文凭和学位文凭之间，还有大学预科文凭(International Foundation Diploma, IFD)——部分大学设有该课程，以及高级国家文凭(Higher National Diploma, HND)——相当于中国的大专文凭。——译者注

$RET_i = 1$,如果这个人退休了,否则 $RET_i = 0$;

$INAC_i = 1$,如果这个人在经济上不活跃①,否则 $INAC_i = 0$;

$UE_i = 1$,如果这个人失业,否则 $UE_i = 0$;

$HNUM_i = 1$,如果这个家庭的人数大于或等于 6,否则 $HNUM_i = 0$;

$SNPAR_i = 1$,如果这个人是单身父亲/母亲,否则 $SNPAR_i = 0$;

$AREA_i = 1$,如果这个人居住在北爱尔兰的 a 地区,否则 $AREA_i = 0$;在北爱尔兰 1991 年的普查中一共有 19 个这样的地区。[21]

除了这些变量外,剥夺程度还可能取决于个人的环境,因为在北爱尔兰,天主教徒是一个相对来说处于劣势的群体,所以也还可能取决于他/她的宗教信仰。(在样本的 13164 个人中,7243 个是男性,5921 个是女性,4364 个是天主教徒,8889 个是新教徒[22])。为了说明这点,还应考虑另外两个变量:

$SEX_i = 1$, 如果这个人是女性,否则 $SEX_i = 0$;

$CT_i = 1$, 如果这个人是天主教徒,否则 $CT_i = 0$。

因此,在这个应用情况下,方程 2.1 设定为:

$$\begin{aligned} D_i = \beta_1 &+ \beta_2 * SEX_i + \beta_3 * CT_i + \beta_4 * AGE_i + \beta_5 * AGE_i^2 + \beta_6 * HIGHED_i \\ &+ \beta_7 * MIDED_i + \beta_8 * RET_i + \beta_9 * INAC_i + \beta_{10} * UE_i \\ &+ \beta_{11} * HNUM_i + \beta_{12} * SNPAR_i + \beta_{13} * AREA_{2i} \cdots \\ &+ \beta_{21} * AREA_{10i} + \varepsilon_i = Z_i + \varepsilon_i \end{aligned} \qquad [2.16]$$

年龄变量的平方(上面的 AGE_i^2)引入了年龄的非线性作用:年龄的增加对 D_i 的边际作用取决于这个增加是在哪个年龄发生的。如果 $\beta_4 < 0$ 且 $\beta_5 > 0$,那么一个人年龄的增加降低了他/她的剥夺分数,但是这个人年纪越大这种影响就越小。[23]

方程统计原理

估计参数 $\hat{\beta}_k$、$\hat{\delta}_1$ 和 $\hat{\delta}_2$ 使观察到样本的可能性最大化,其中样本 $N_1 =$

---

① 即不从事经济活动。——译者注

6042，$N_2 = 4594$，$N_3 = 2528$（见方程 2.4）。

这些估计值列于对应次序 logit 模型的表 2.1 以及对应次序 probit 模型的表 2.2 中。表 2.1 和表 2.2 中的 z-ratio 是估计系数与他们的估计标准误差之间的比：这些 z-ratio（渐近）服从零假设为相关系数等于零的 $N(0, 1)$分布。[24]

表 2.1 北爱尔兰社会剥夺程度的次序 logit 模型结果

Ordered Logit Estimates; *Log Likelihood* = −12423.562; *Number of obs* = 13164;
LR $\chi^2(20)$ = 2571.98; *Prob*>χ^2 = 0.0000; *Pseudo* R^2 = 0.0938

y	Coefficient	Standard Error	z	P>|z|	[95% Conf.	Interval]
sex	−0.1649064	0.0360722	−4.572	0.000	−0.2356066	−0.0942061
ct	0.1829855	0.0393496	4.650	0.000	0.1058617	0.2601092
age	−0.0253924	0.0049969	−5.082	0.000	−0.0351862	−0.0155987
age2	0.0008173	0.0000946	8.638	0.000	0.0006319	0.0010028
ret	0.6264438	0.0875236	7.157	0.000	0.4549007	0.797987
inac	1.297307	0.0765827	16.940	0.000	1.147207	1.447406
ue	1.460677	0.0619881	23.564	0.000	1.339182	1.582171
highed	−0.9081216	0.0683311	−13.290	0.000	−1.042048	−0.774195
mided	−0.6692574	0.1230722	−5.438	0.000	−0.9104745	−0.4280403
hnum	0.9214472	0.0551311	16.714	0.000	0.8133923	1.029502
snpar	0.305537	0.0637575	4.792	0.000	0.1805746	0.4304994
ard	−0.4497501	0.0642359	−7.002	0.000	−0.5756502	−0.3238499
dwn	0.0492598	0.0680278	0.724	0.469	−0.0840723	0.1825919
crk	−0.3161319	0.0706348	−4.476	0.000	−0.4545735	−0.1776903
ant	0.3472878	0.0717874	4.838	0.000	0.206587	0.4879887
col	0.4384781	0.0715933	6.125	0.000	0.2981578	0.5787985
arm	0.3824736	0.0740609	5.164	0.000	0.237317	0.5276303
ban	0.2364069	0.066677	3.546	0.000	0.1057223	0.3670914
dry	0.0791624	0.0787678	1.005	0.315	−0.0752198	0.2335445
frm	0.6350174	0.0747171	8.499	0.000	0.4885745	0.7814602
_cut1	0.1804476	0.074681		(Ancillary parameters)		
_cut2	2.058464	0.0772157				

y	Probability	Observed
1	Pr($xb+u<$ _cut1)	0.4591
2	Pr(_cut1 $< xb+u <$ _cut2)	0.3493
3	Pr(_cut2 $< xb+u$)	0.1917

表 2.2 北爱尔兰社会剥夺程度的次序 Probit 模型结果

Ordered Probit Estimates; *Log Likelihood* = −12426.255; *Number of obs* = 13164;
LR $\chi^2(20)$ = 2566.60; *Prob* > χ^2 = 0.0000; *Pseudo* R^2 = 0.0936

y	*Coefficient*	*Standard Error*	*z*	*P* > |*z*|	[95% *Conf.*	*Interval*]
sex	−0.1003127	0.021542	−4.657	0.000	−0.1425343	−0.0580911
ct	0.112619	0.0235271	4.787	0.000	0.0665067	0.1587314
age	−0.0154206	0.002965	−5.201	0.000	−0.021232	−0.0096093
age2	0.0004965	0.000056	8.872	0.000	0.0003868	0.0006062
ret	0.3679154	0.0528714	6.959	0.000	0.2642894	0.4715415
inac	0.7680422	0.0445939	17.223	0.000	0.6806398	0.8554446
ue	0.8555121	0.0357888	23.904	0.000	0.7853673	0.925657
highed	−0.5562051	0.0401161	−13.865	0.000	−0.6348313	−0.4775789
mided	−0.4048568	0.0721096	−5.614	0.000	−0.546189	−0.2635246
hnum	0.5459127	0.0328011	16.643	0.000	0.4816237	0.6102017
snpar	0.1748777	0.0379915	4.603	0.000	0.1004157	0.2493396
ard	−0.2729771	0.03806	7.172	0.000	−0.3475734	−0.1983808
dwn	0.0216987	0.0403927	0.537	0.591	−0.0574696	0.100867
crk	−0.1831732	0.0417875	4.383	0.000	−0.2650752	−0.1012712
ant	0.2026288	0.0428712	4.726	0.000	0.1186028	0.2866547
col	0.2518903	0.0428189	5.883	0.000	0.1679668	0.3358137
arm	0.2122576	0.0444114	4.779	0.000	0.1252129	0.2993024
ban	0.129325	0.0399411	3.238	0.001	0.0510418	0.2076082
dry	0.0360175	0.0471155	0.764	0.445	−0.0563271	0.1283622
frm	0.3626456	0.0447755	8.099	0.000	0.2748871	0.450404
_cut1	0.1009536	0.0444785			(Ancillary parameters)	
_cut2	1.212453	0.0453671				

y	*Probability*	*Observed*
1	Pr(*xb* + *u* < _cut1)	0.4591
2	Pr(_cut1 < *xb* + *u* < _cut2)	0.3493
3	Pr(_cut2 < *xb* + *u*)	0.1917

格林(2000:831—833)提供了一些测量离散因变量方程的"拟合优度"的建议。他提议,至少我们应该报告对数似然函数最大化了的值。L_1 的值就是最大对数似然值,列于表格的顶部(分别为−12423.56 和−12426.25)。由于方程中所有斜率都为零的假设经常令人很感兴趣,所以还应该报告"完整"模型和"只

有截距项"模型的比较结果。表 2.1 和表 2.2 顶部的 χ^2 值(分别为 2571.98 和 2566.6)等于 $2(L_1 - L_0)$，其中 L_0 是当解释变量只有常数项时的对数似然函数的值，L_1 则如前面观察到的，是当所有的解释变量都包括在内时对数似然函数的值。自由度等于被估计的斜率系数的个数。这些 χ^2 的值完全拒绝了这个模型不比"只有截距项"的模型更具解释力的零假设。[25]

"pseudo-R^2"被定义为 $1 - L_1/L_0$，这归功于麦克法登(McFadden, 1973)。这个值的范围固定在 0 与 1 之间。值为 0 对应于所有的斜率系数均为 0，值为 1 对应于理想预测(也就是，$L_1 = 0$)。遗憾的是，正如格林(2000)指出的，介于 0 和 1 之间的值并没有自然的解释，虽然提示是 pseudo-R^2 的值随着模型拟合的改善而增加。还有人提出其他方法。本阿齐瓦和勒曼(Ben-Akiva & Lerman, 1985)以及凯和利特尔(Kay & Little, 1986)提出了一种通过预测规则测量正确预测的平均概率的拟合方法。卡默(Carmer, 1999)提出了一种纠正本阿齐瓦/勒曼测量不足之处的方法，考虑了在不平衡样本中，对频数少的结果进行的预测往往非常糟糕这样一个事实。维尔和齐默尔曼(Veall and Zimmerman)(1996)在他们的 pseudo-R^2 测量中，主张在多类别 probit 或者 logit 类型的模型中，只有麦克法登(1973)的方法"看起来有价值"。

另外一种拟合优度的"点"测量方法可能是评估模型的预测能力。对在一种预测规则的基础上，这些评估在二分选择模型中是例行程序(比如说 $Y_i = 1$ 如果 $\hat{P}_i > 0.5$，否则 $\hat{P}_i = 0$)，预测到的命中($Y_i = 1$)和不命中($Y_i = 0$)与真实的命中与不命中进行比较。这个方法可以扩展到多元结果模型，其预测可以基于如果 $\hat{P}_{im} = \mathrm{Max}_j(\hat{P}_{ij})$，则 $Y_i = m$ 的规则。然后这些预测可以与估计所有情况都处于因变量最常见类别中的这样一种"简单"模型进行比较，从这个简单模型到完整模型减少了的误差的百分比就可以计算出来了。但是，这种方法并不是没有缺陷。首先，与线性回归模型中系数被选择以最大化 R^2 的情况不同，在离散选择模型中，系数的估计值并不最大化任何拟合优度测量。所以，不管拟合优度怎么被测量，以拟合优度为基础对模型进行评估都可能是误导性的。第二，这些预测严重取决于采取的预测规则，但可能这些预测规则本来非

常不适合样本的需要。比如，在一个二分模型中，如果样本是不平衡的——就是说有更多的1,更少的0——则模型应该预测估计概率最大的结果这样的一个规则可能永远无法预测到一个1(或者0)。

估计值

表2.1和表2.2中的所有系数,除了两个与居住在County Down和County Derry地区(表中的*dwn*和*dry*变量)相关的系数之外,其他都明显不等于0。在表2.1的“$P>|z|$”这列下面的数字表示,在*dwn*和*dry*的系数等于0的零假设下,分别有46.6%和31.5%的机会,会观察到处于分布的尾部超出观测值0.0492598和0.0791624的值。在表2.2的“$P>|z|$”这列下面的数字表示,在这两个系数等于0的零假设下,观察到一个超出观测值的数值的机会为“零”。在“95% Conf. Interval”这列下面的数字表示如果它们的相关系数要在5%显著性水平下不等于零,估计值必须超出的上下限。在表2.1和表2.1中,对*dwn*和*dry*系数的估计值处于它们的95%区间内,而其他系数的估计值却在它们的区间之外。

在解释单个系数的时候,必须强调其他条件不变(“其他事情相同”)的重要性。变量*ct*的估计值是正的,表示在其他条件不变的情况下,天主教徒与新教徒相比,有更高的被严重剥夺的概率和更低的没有被剥夺的概率。这并不是说每个天主教徒都比每个新教徒有更高的被严重剥夺的概率。相反,正确的说法应该是,给定两个除了在宗教信仰之外其他每种特征都相似的人,天主教徒比新教徒更加可能被严重剥夺,更加不可能没有被剥夺。变量*sex*的估计值是负的,表示其他条件相同的女性比男性有更低的被严重剥夺的概率和更高的没有被剥夺的概率。其他系数的解释相似。比如,变量*highed*和*mided*的负的估计值表示,在其他条件相同的情况下,具有高学历文凭的人比没有高学历文凭的人有更低的被严重剥夺的概率和更高的没有被剥夺的概率。正如前面的讨论所说明的,系数估计值的方向只允许对处于两端的结果,其概率随着相关变量值的改变而变化的方向进行估计。处于中间的结果其概率的变化方向不能推断。比如,从对估计值的审查中,我们无法判断女性相对于男性来说被轻度

剥夺的概率更大还是更小。

临界点

估计出来的临界值呈现于表中各估计值的下面。表 2.1 和表 2.2 的下方显示的部分表示各类别的概率如何从拟合方程中计算出来。[26]这部分显示的结果通过在 oprobit 和 ologit 命令语法中加入“table”这个选择命令而得，这些命令列于第 4 章的程序表中。有序回归背后的假设是观察到的类别代表了粗略的但以一种隐含的等级正确排序的差别。属于这些类别的概率的确定基于隐含的潜在变量的值跨越特定界限的概率，这些界限是通过临界点的值而建立的。

在次序 logit 模型中，这两个临界点被估计为[27] $_cut1 = 0.1804476$ 和 $_cut2 = 2.058464$，在方程 2.5a 至 2.5c 的符号中，$\hat{\delta}_1 = _cut1$，$\hat{\delta}_2 = _cut2$。因此，在次序 logit 模型中，一个人没有被剥夺的概率是 $\Pr(\hat{Z}_i + \varepsilon_i \leqslant 0.1804476)$，被轻度剥夺的概率为 $\Pr(0.1804476 \leqslant \hat{Z}_i + \varepsilon_i \leqslant 2.058464)$，被严重剥夺的概率为 $\Pr(\hat{Z}_i + \varepsilon_i \geqslant 2.058464)$。在次序 probit 模型中，这两个临界点被估计为 $_cut1 = 0.1009536$ 和 $_cut2 = 1.212453$。在这个模型中，一个人没有被剥夺的概率是 $\Pr(\hat{Z}_i + \varepsilon_i \leqslant 0.1009536)$，被轻度剥夺的概率为 $0.1009536 \leqslant \hat{Z}_i + \varepsilon_i \leqslant 1.212453)$，被严重剥夺的概率为 $\Pr(\hat{Z}_i + \varepsilon_i \geqslant 1.212\ 453)$。这两个模型中的临界点不一样的原因是斜率系数的估计值也完全不同。但是，每个模型中估计出来的斜率和临界系数的结合使从这两个模型中得出的预测非常相似。

概率估计值：个人的估计

用估计到的 $\hat{Z}_i$——记住 $\hat{Z}_i = \sum_{k=1}^{k} \hat{\beta}_k X_{ik}$，可利用表 2.1 中的估计值，结合每个人的决定变量的取值而得出——STATA 会对样本 13164 个人中的每个人，通过分别计算方程 2.5a，2.5b 和 2.5c 中的 $\hat{p}_{i1}$，$\hat{p}_{i2}$和 $\hat{p}_{i3}$，对他们处于三种不同的剥夺程度的概率进行预测。表 2.3 显示了通过误差项 ε_i 的逻辑分布对样本前 25 个人的估计结果。从表 2.3 我们可以看到，给定他/她的情况，第 17 个人有非常低的被严重剥夺的概率（4%）和非常高的没有被剥夺的概率

(78%)。另一方面,第二个人的情况表示他/她有非常高的被严重剥夺的概率(69%)和非常低的没有被剥夺的概率(6%)。把 logit 模型预测的概率与假设 ε_i 为正态分布而得出的概率进行比较是非常有启发性的。表 2.4 列出了后者的估计结果。对表 2.3 和表 2.4 进行的比较显示,尽管系数估计值在 logit 和 probit 模型之间存在差异,预测到的概率却非常相似。

表 2.3 处于不同剥夺水平的次序 Logit 概率计算值:样本前 25 个观测值

P_{num}	$p1$	$p2$	$p3$
1	0.2468799	0.4350583	0.3180617
2	0.0635542	0.2438714	0.6925744
3	0.1456977	0.3815914	0.472711
4	0.221723	0.4290327	0.3492444
5	0.1984064	0.4197503	0.3818433
6	0.0774309	0.2769676	0.6456016
7	0.3072922	0.4363922	0.2563156
8	0.6020886	0.3061397	0.0917717
9	0.6067917	0.303063	0.0901453
10	0.5668221	0.3285579	0.10462
11	0.3003936	0.4370245	0.2625819
12	0.4497833	0.3926532	0.1575635
13	0.1598765	0.3946234	0.4455001
14	0.5337384	0.348435	0.1178265
15	0.6265731	0.2899149	0.0835119
16	0.5928456	0.3121286	0.0950258
17	0.7833934	0.1760466	0.04056
18	0.5323462	0.3492447	0.1184092
19	0.5169058	0.3580674	0.1250268
20	0.6024157	0.3059263	0.091658
21	0.5469929	0.3406158	0.1123913
22	0.5918246	0.3127853	0.0953901
23	0.7823876	0.1768215	0.0407909
24	0.5389312	0.3453955	0.1156733
25	0.4514354	0.3918849	0.1566797

注:p_{num}= 个人序号。
$p1$=“没有被剥夺”的概率。
$p2$=“轻度被剥夺”的概率。
$p3$=“严重被剥夺”的概率。

表 2.4 处于不同剥夺水平的次序 Probit 概率计算值:样本前 25 个观测值

P_{num}	$p1$	$p2$	$p3$
1	0.2564767	0.4197788	0.3237445
2	0.0561468	0.2607234	0.6831297
3	0.148693	0.3789889	0.4723181
4	0.2297609	0.4152423	0.3549968
5	0.2077654	0.4091158	0.3831188
6	0.0723101	0.2918726	0.6358173
7	0.3164209	0.4204631	0.263116
8	0.5960203	0.3162006	0.0877791
9	0.6006402	0.3134694	0.0858905
10	0.5609338	0.3361012	0.102965
11	0.3078035	0.4210712	0.2711253
12	0.4515077	0.3873188	0.1611736
13	0.1692549	0.3920921	0.438653
14	0.5286788	0.3530054	0.1183157
15	0.6201291	0.3016761	0.0781948
16	0.5868787	0.3215305	0.0915909
17	0.7813467	0.189157	0.0294963
18	0.5321902	0.3512332	0.1165766
19	0.5122991	0.3610432	0.1266577
20	0.5962939	0.3160395	0.0876666
21	0.5415704	0.3464162	0.1120134
22	0.5858997	0.3220953	0.092005
23	0.7803024	0.1899634	0.0297342
24	0.5342544	0.3501835	0.1155621
25	0.449369	0.3881331	0.1624979

注:p_{num}=个人序号。
$p1$=“没有被剥夺”的概率。
$p2$=“轻度被剥夺”的概率。
$p3$=“严重被剥夺”的概率。

对列出的 25 个人中的每个人来说,用 logit 和 probit 方法估计出的处于三种剥夺程度的概率并没有很大不同。从某种意义上说,用哪个模型并不重要,预测结果非常相似。

个人概率可用于生成剥夺程度的样本统计值。这些统计值列于表 2.5

(logit 模型)和表 2.6(probit 模型)中。

这两个表中显示的 $\hat{p}_{i1}$，$\hat{p}_{i2}$和 $\hat{p}_{i3}$的均值(分别用 $\bar{p}_1$，$\bar{p}_2$ 和 $\bar{p}_3$ 表示)在 logit 和 probit 模型间稍有不同。但是，取整数之后，这些不同消失了且估计值分别为：46%，35%和 19%。在 logit 模型中，用上述方法算出的平均概率，等于在三个剥夺类别中的样本比例。[28]这是次序 logit 模型的一个特性。在 probit 模型下平均概率很接近，但不等于样本比例，这种情况也是在实践中经常见到的。

表 2.5 和表 2.6 中的平均概率总和精确地为 1。这是因为个人的概率估计(平均概率从这些概率估计中计算而得)总和为 1。

$\hat{p}_{i1}$，$\hat{p}_{i2}$和 $\hat{p}_{i3}$的中位数(第 50 个百分位数)，如表 2.5 和表 2.6 所示，在 logit 和 probit 模型间存在些微的差异，但是这些差异再次在取整数之后消失了，产生的中位数概率的估计值分别为 48%，37%和 14%。对这两个模型中的任何一个来说，中位数概率都不一定总和为 1，在这个应用实例中，它们的总和确实并不等于 1。原因在于虽然对任何一个人估计值的总和为 1，但是 $\hat{p}_{i1}$，$\hat{p}_{i2}$和 $\hat{p}_{i3}$的中位数却不一定与同一个人相关。比如，如果 $\hat{p}_{i1}$，$\hat{p}_{i2}$和 $\hat{p}_{i3}$($i=1, \cdots, 13164$)的值按照递增或递减的顺序排列，那么在 $\hat{p}_{i1}$排列中处于第 50 个百分位数的那个人可能与在 $\hat{p}_{i3}$中第 50 个百分位数的那个人不同。在 Logit 的估计值方面(表 2.5)，没有被剥夺的概率($\hat{p}_{i1}$)处于最低的百分位数的是 6.7%，在这个等级里面最低的估计值为 1.9%。在这个量表的另一端，$\hat{p}_{i1}$在最高的百分位数的值为 83.1%，在这个级别里面最高的估计值为 87%。

表 2.5 不同剥夺程度的样本统计值：Logit 模型

	Pr(y = 1)			
	Percentiles	Smallest		
1%	0.0675202	0.0185615		
5%	0.1466481	0.0198162		
10%	0.1875675	0.0300967	Obs	13164
25%	0.303902	0.0347162	Sum of Wgt.	13164
50%	0.4843619		Mean	0.4592781

（续表）

		Largest	Standard Deviation	0.1889648
75%	0.5993176	0.8697239		
90%	0.6954355	0.8697425	Variance	0.0357077
95%	0.7368268	0.8699151	Skewness	−0.1866478
99%	0.8308569	0.8699151	Kurtosis	2.155329

Pr($y=2$)

	Percentiles	Smallest		
1%	0.1389147	0.0915195		
5%	0.2075179	0.0969696		
10%	0.239251	0.1077327	Obs	13164
25%	0.2989144	0.1077327	Sum of Wgt.	13164
50%	0.3674155		Mean	0.3493639
		Largest	Standard Deviation	0.076207
75%	0.4150781	0.4377985		
90%	0.4339717	0.4377985	Variance	0.0058075
95%	0.4366746	0.4377985	Skewness	−0.8337451
99%	0.4377566	0.4377985	Kurtosis	2.917122

Pr($y=3$)

	Percentiles	Smallest		
1%	0.0301859	0.0223522		
5%	0.0517813	0.0223522		
10%	0.0627571	0.0223856	Obs	13164
25%	0.0927392	0.0223891	Sum of Wgt.	13164
50%	0.1399815		Mean	0.191358
		Largest	Standard Deviation	0.1405856
75%	0.2593728	0.8095671		
90%	0.3984033	0.8312855	Variance	0.0197643
95%	0.4708132	0.8832142	Skewness	1.451671
99%	0.6786129	0.889919	Kurtosis	4.969932

表 2.6 不同剥夺程度的样本统计值:Probit 模型

		Pr(y = 1)		
	Percentiles	Smallest		
1%	0.0610129	0.0089802		
5%	0.1495682	0.0095785		
10%	0.1936998	0.0190097	Obs	13164
25%	0.311691	0.0255515	Sum of Wgt.	13164
50%	0.4870896		Mean	0.4602124
		Largest	Standard Deviation	0.1855043
75%	0.5928672	0.8747257		
90%	0.6891779	0.8747432	Variance	0.0344118
95%	0.7340055	0.8749362	Skewness	−0.212917
99%	0.8299778	0.8749362	Kurtosis	2.250704
		Pr(y = 2)		
	Percentiles	Smallest		
1%	0.1500103	0.0957713		
5%	0.2209043	0.09959		
10%	0.2547602	0.1132009	Obs	13164
25%	0.3088254	0.1132009	Sum of Wgt.	13164
50%	0.3662931		Mean	0.3488479
		Largest	Standard Deviation	0.0663654
75%	0.4044376	0.421618		
90%	0.418923	0.421618	Variance	0.0044044
95%	0.4207648	0.421618	Skewness	−1.039971
99%	0.421588	0.421618	Kurtosis	3.528536
		Pr(y = 3)		
	Percentiles	Smallest		
1%	0.0194342	0.011863		
5%	0.0412402	0.011863		
10%	0.0542446	0.011892	Obs	13164
25%	0.0890826	0.0118946	Sum of Wgt.	13164
50%	0.1402645		Mean	0.1909397
		Largest	Standard Deviation	0.1432815
75%	0.2674853	0.7992974		
90%	0.4023936	0.8322631	Variance	0.0205296
95%	0.4708187	0.8908314	Skewness	1.315678
99%	0.668156	0.8952485	Kurtosis	4.483449

概率估计值:均值的估计

在前文中,平均概率的估计值从个人概率 $\hat{p}_{i1}$, $\hat{p}_{i2}$ 和 $\hat{p}_{i3}$ ($i = 1, \cdots, 13164$)的估计值中计算得出。在 logit 模型中平均概率正好等于样本比例,在 probit 模型中,他们也很接近样本比例。但是,还有另外一种方法计算平均概率(见格林, 2000: 879),且这种方法产生的结果与上面的不同。设 $\bar{X}_k = \sum_{i=1}^{N} X_{ik}/N$ 为第 k 个决定变量的平均值,让 $\hat{Z}_i$ 的均值等于:

$$\bar{Z} = \sum_{i=1}^{N} \hat{Z}_i/N = \sum_{i=1}^{N}\Big(\sum_{k=1}^{K} \hat{\beta}_k X_{ik}\Big)/N$$
$$= \sum_{k} \hat{\beta}_k \sum_{i} X_{ik}/N = \sum_{k} \hat{\beta}_k \bar{X}_k \qquad [2.17]$$

然后计算 $\hat{p}_1$, $\hat{p}_2$,和 $\hat{p}_3$ 如下:

$$\hat{p}_1 = \Pr(\varepsilon_i \leqslant \hat{\delta}_1 - \bar{Z}) \qquad [2.18a]$$

$$\hat{p}_2 = \Pr(\hat{\delta}_1 - \bar{Z} < \varepsilon_i \leqslant \hat{\delta}_2 - \bar{Z}) \qquad [2.18b]$$

$$\hat{p}_3 = \Pr(\varepsilon_i \geqslant \hat{\delta}_2 - \bar{Z}) \qquad [2.18c]$$

通常,$\bar{p}_1 \neq \hat{p}_1$,$\bar{p}_2 \neq \hat{p}_2$,$\bar{p}_3 \neq \hat{p}_3$,这并不奇怪。因为尽管 $\bar{p}_j$ 和 $\hat{p}_j$($j = 1, 2, 3$)两者都旨在测量处于不同剥夺水平的总体概率,但是它们的计算方式却非常不同。$\bar{p}_j$ 计算为估计到的个人概率 $\hat{p}_{ij}$ 的均值,而 $\hat{p}_{ij}$ 是从个人决定变量的值结合估计到的临界点计算出来的。另一方面,$\hat{p}_j$ 绕过了个人概率,反而是用个人 Z_i 的均值 $\bar{Z}$(或者,等于用决定变量 X_{ik} 的均值 $\bar{X}_k$ 计算而得),结合估计到的临界点而得出。表 2.7 和表 2.8 比较了 logit 模型中 $\bar{p}_j$ 的值和 probit 模型中 $\hat{p}_j$ 的值。

表 2.7 处于不同剥夺程度的总体概率:logit 模型

概　　率	计算为个人概率的均值: $\bar{p}$	从决定变量的均值计算而得: $\hat{p}$
没有被剥夺	0.459	0.447
轻度被剥夺	0.349	0.394
严重被剥夺	0.191	0.159

表 2.8　处于不同剥夺程度的总体概率：probit 模型

概　　率	计算为个人概率的均值：$\bar{p}$	从决定变量的均值计算而得：$\hat{p}$
没有被剥夺	0.460	0.452
轻度被剥夺	0.349	0.387
严重被剥夺	0.191	0.161

边际概率的计算：连续变量

模型中唯一的连续变量，如方程 2.16 所示，是 AGE_i 和 AGE_i^2。为了计算个人年龄每增加 1 年对他/她处于三种剥夺程度概率的影响，我们需要在相关点上估计合适的密度函数并乘以与两个变量相关的系数估计值（见方程 2.12a 至 2.12c，以及 2.13a 至 2.13c）。

还可以通过下列两个方式中的任何一种完成：

1. 对样本中的每个人，在相关点上对其计算边际效应。对个人 $i(i=1,\cdots,13164)$，这些相关点为：$\hat{\delta}_1-\hat{Z}_i$ 对应于 $\Pr(Y_i=1)$的边际效应；$\hat{\delta}_2-\hat{Z}_i$ 和 $\hat{\delta}_1-\hat{Z}_i$ 对应于 $\Pr(Y_i=2)$的边际效应；$\hat{\delta}_2-\hat{Z}_i$ 对应于 $\Pr(Y_i=3)$的边际效应。用这些估计值乘以 AGE_i 的系数估计值（$\hat{\beta}_4$），同时也用这些估计值乘以 AGE_i^2 的系数估计值（$\hat{\beta}_5$）。这两个结果相加就得到了每个人年龄的一个小变化（1 年）对他/她的概率的边际效应。然后所有这些个人作用的均值就产生了样本中每个人的年龄增加 1 年对（处于不同的剥夺程度的）概率的平均影响。

2. 计算样本中所有个人 $\hat{Z}_i$ 值的均值（见方程 2.17）。如果这个均值记为 $\bar{Z}$，在相关点计算边际效应：$\hat{\delta}_1-\bar{Z}$ 对应于 $\Pr(Y_i=1)$的边际效应；$\hat{\delta}_2-\bar{Z}$ 和 $\hat{\delta}_1-\bar{Z}$ 对应于 $\Pr(Y_i=2)$的边际效应；$\hat{\delta}_2-\bar{Z}$ 对应于 $\Pr(Y_i=3)$的边际效应。

用这些估计值乘以 AGE_i 的系数估计值，同时也用这些估计值乘以 AGE_i^2 的系数估计值。这两个结果相加就得到了每个人年龄的一个小变化（1 年）对他/她的概率的边际效应。然后所有这些个人作用的均值就产生了样本中所有人的平均年龄增加 1 年对（处于不同的剥夺程度的）概率的影响。

表 2.9 和表 2.10 展示了用以上两种不同的方法计算出来的 logit 和 probit

模型中年龄增加的边际效应。

表 2.9 年龄对处于不同剥夺程度的边际效应:logit 模型

概　　率	计算为个人边际效应的均值	从决定变量的均值计算而得
没有被剥夺	−0.00523	−0.00607
轻度被剥夺	0.00191	0.00279
严重被剥夺	0.00332	0.00329

表 2.10 年龄对处于不同剥夺程度的边际效应:probit 模型

概　　率	计算为个人边际效应的均值	从决定变量的均值计算而得
没有被剥夺	−0.00524	−0.00591
轻度被剥夺	0.00171	0.00226
严重被剥夺	0.00353	0.00365

边际概率的计算:虚拟变量

在前面我们观察到一个虚拟变量的作用应该通过比较它取其中一个值时的概率与取另一个值的概率进行分析,在这个比较过程中其他变量的值保持不变。这个方法现在用于分析宗教信仰对处于不同剥夺程度的概率的影响,通过比较(对所有的 i)$CT_i=1$(宗教虚拟变量)的情况与(对所有的 i)$CT_i=0$ 的情况。与连续变量的情形一样,这种广义的方法可以采用以下两种途径中的任意一种:

1. 假设其他条件均同且样本中的所有人都是天主教徒 $CT_i=1(i=1\cdots 13164)$。那么在这种情况下,$\hat{Z}_i$(Z_i 的估计值)从公式 2.16 中用系数估计值以及对所有 $i CT_i=1$ 计算而得。把这个估计出来的值叫做 $\hat{Z}_i^c$。

让 $\hat{p}_{ij}^c$ 表示在这个假设的情况下个人 i 处于剥夺程度 $j(j=1, 2, 3)$的(计算出来的)概率,其中这些概率从方程 2.5a 至 2.5c 中——或者,如果使用 probit 模型,从方程 2.11a 至 2.11c 中——用 $\hat{Z}_i=\hat{Z}_i^c$ 计算而得。现在假设其他条件相同且样本中的所有人都是新教徒,对所有 $i=1\cdots 13164$, $CT_i=0$,让 $\hat{Z}_i^p$ 表示 Z_i 的估计值。从方程 2.6, $\hat{Z}_i^c=\hat{Z}_i^p+\hat{\beta}_3$。让 $\hat{p}_{ij}^p$ 表示在这些假设的情况

下个人 i 处于剥夺程度 $j(j=1, 2, 3)$ 的(计算出来的)概率。对任何样本中的个人 i, $\hat{p}^c_{ij}$ 和 $\hat{p}^p_{ij}$ 的差别完全来自于宗教信仰的差别：$\hat{p}^c_{ij}$ 是用 $\hat{Z}^c_i$——也就是，天主教徒这个变量等于 1 时——计算出来的；$\hat{p}^p_{ij}$ 是用 $\hat{Z}^p_i$，天主教徒这个变量等于 0 时计算出来的；在其他变量的值没有任何变化的情况下。如果 $\bar{p}^c_j=\sum_{i=1}^{N}\hat{p}^c_{ij}/N$ 和 $\bar{p}^p_j=\sum_{i=1}^{N}\hat{p}^p_{ij}/N$ 分别是两组概率估计值的均值，那么它们之间的差异就测量出了宗教信仰对处于不同剥夺程度的平均概率的影响。

2. 另一个方法是比较对样本中的所有人，变量 CT_i 取两个不同值时的两种概率，且在每种情况下其他变量都取样本均值。让 $\bar{Z}^c$ 和 $\bar{Z}^p$ 分别表示对所有 $i=1\cdots13164$, $CT_i=1$ 和 $CT_i=0$ 时 $\bar{Z}$ 的值，且在每一种情况下其他变量 $X_{ik}=\bar{X}_k$。由公式 2.16，$\bar{Z}^c=\bar{Z}^p+\hat{\beta}_3$。实际上，这个应用除了宗教信仰外还包括了样本的平均特征，而且在一种情况下它是天主教徒，在另一种情况下它是新教徒。那么，在方程 2.5a 至 2.5c 中使用 $\bar{Z}^c$，然后 $\bar{Z}^p$，就可以对这两种假定的情形(前一种情形下这个人为天主教徒，后者这个人为新教徒)计算出(处于没有被剥夺，轻度被剥夺，和严重被剥夺的)三种概率的估计值。如果这两种概率分别记为 $\hat{q}^c_j$ 和 $\hat{q}^p_j(j=1, 2, 3)$，那么它们之间的差异就测量出了宗教信仰对处于不同剥夺程度的概率的影响。

表 2.11 展示了在 logit 模型下 $\bar{p}^c_j$ 和 $\bar{p}^p_j$，以及 $\hat{q}^c_j$ 和 $\hat{q}^p_j(j=1, 2, 3)$ 的值，表 2.12 则表示 probit 模型下的情况。这两个表有两个重要的特点。首先，对于两种计算边际效应的方法中的任何一种，logit 和 probit 概率之间都几乎没有差异。其次，对任何一个模型来说，用两种不同的方法计算出来的概率之间存在着非常大的差异。注意，用个人概率的均值计算出来的处于严重被剥夺的概率显著地高于用样本的平均特征计算出来的处于严重被剥夺的概率。

表 2.11 宗教信仰对处于不同剥夺程度的概率的影响：logit 模型

概率	计算为个人边际效应的均值		从决定变量的均值计算而得	
	CAT	*PRT*	*CAT*	*PRT*
没有被剥夺	0.432	0.472	0.417	0.462
轻度被剥夺	0.360	0.346	0.407	0.387
严重被剥夺	0.207	0.182	0.176	0.151

表 2.12 宗教信仰对处于不同剥夺程度的概率的影响:probit 模型

概 率	计算为个人边际效应的均值		从决定变量的均值计算而得	
	CAT	*PRT*	*CAT*	*PRT*
没有被剥夺	0.433	0.473	0.421	0.466
轻度被剥夺	0.359	0.346	0.398	0.382
严重被剥夺	0.208	0.191	0.181	0.152

处于严重被剥夺的状态是拥有剥夺决定变量的"极端"值的结果。极端值的影响在个人其他变量值被设置为样本均值时被减弱了。从另一方面来说,在概率计算中使用个人的变量值时,极端值被允许充分地发挥了作用。

那么哪一种方法是适合使用的呢?关键的问题是,当我们感兴趣的虚拟变量取各不相同的两种值的时候,其他变量的值是如何保持不变的。实际上,第二种方法对每个人都分配了样本的平均值。但是,平均值并不是神圣不可侵犯的,这个分配的共同值也可以用中位数或者任意其他值,且这些不同分配方式中的每一种都会在计算出来的概率 $\hat{q}_j$ 上产生一个不同的结果。

第一种方法就没有这个缺陷。因为个人的值没有被干扰,计算出来的个人的概率有一个独特的结果,所以在两种不同的情形下,平均[29]概率 $\bar{p}_j$ 也有一个独特的结果。

对次样本的估计:特征与系数

宗教信仰由于方程 2.16 中 β_3 的存在而对处于不同剥夺程度的概率(表 2.11和 2.12)有影响。这意味着,在第一种方法下 $\hat{Z}_i^c \neq \hat{Z}_i^p$,在第二种方法下 $\bar{Z}^c \neq \bar{Z}^p$。因此,样本中的每个人 $CT_i = 1$ 时被剥夺的概率不同于样本中每个人 $CT_i = 0$ 时的概率。但这导致我们考虑这样一种可能性,即对方程 2.16 中的每个变量,"天主教徒"的系数都不同于"新教徒"的系数。也就是说,方程 2.16的估计应该允许天主教徒与新教徒的系数互不相同。有两种(几乎等同的)的操作方法。第一种只估计一个方程,但允许方程中的每个天主教徒的系

数都与相应的新教徒的系数不同。第二种对天主教徒和新教徒次样本分别估计两个不同的方程。为了执行第一种策略，定义方程如下：

$$\begin{aligned}D_i = \beta_1 &+ \gamma_1 * CT_i + \beta_2 * SEX_i + \gamma_2 * CT_i * SEX_i + \beta_3 * AGE_i \\ &+ \gamma_3 * CT_i * AGE_i + \beta_4 * AGE_i^2 + \gamma_4 * CT_i * AGE_i^2 \\ &+ \beta_5 * HIGHED_i + \gamma_5 * CT_i * HIGHED_i + \beta_6 * MIDED_i \\ &+ \gamma_6 * CT_i * MIDED_i + \beta_7 * RET_i + \gamma_7 * CT_i * RET_i \\ &+ \beta_8 * INAC_i + \gamma_8 * CT_i * INAC_i + \beta_9 * UE_i \\ &+ \gamma_9 * CT_i * UE_i + \beta_{10} * HNUM_i + \gamma_{10} * CT_i * HNUM_i \\ &+ \beta_{11} * SNPAR_i + \gamma_{11} * CT_i * SNPAR_i + \beta_{13} * AREA_{2i} \\ &+ \cdots + \beta_{21} * AREA_{10i} + \gamma_{13} * CT_i * AREA_{2i} + \cdots \\ &+ \gamma_{21} * CT_i * AREA_{10i} + \varepsilon_i = Z_i + Z_i^c + \varepsilon_i \end{aligned} \qquad [2.19]$$

通过在方程中加入“交互变量”——即用方程 2.16 中的原解释变量乘以天主教徒虚拟变量——某个特定变量对剥夺的影响就可依据受关注的人是新教徒还是天主教徒而不同。在方程 2.19 中，β_k 是“新教徒”的系数[30]，而 γ_k——交互变量的系数——代表了作为天主教徒而导致的对这些系数的附加贡献。因此 γ_k 的值测量了“交互作用”的强度。所以对 $\gamma_k = 0$ 的检验就是对第 k 个解释变量与个人的宗教信仰之间没有交互影响这个零假设的检验。换一种说法，$\gamma_k = 0$ 意味着在第 k 个解释变量上新教徒与天主教徒的系数之间没有差别。与两个不同方程的方法相比，这个单一方程“整合”方法的一大优势在于，它使轻易地检验某个变量对这两组人是否有不同的影响成为可能。实际上，随后的内容还对这个单一方程“整合”方法进行了一定程度的探讨。但是，为了给第二种方法增添风趣，方程 2.16 对天主教徒和新教徒次样本分别进行了估计。表2.13 和表 2.14 分别展示了当误差被假设为逻辑分布时，方程 2.16 对天主教徒和新教徒的估计结果。[31]

在表 2.13 和表 2.14 所示的方程估计值中，变量 *dun* 和 *dry* 被丢掉了。这是因为它们的系数单个地（在 Z-score 上）以及共同地（在似然比检验上）都不显

著地不等于 0。这是一个方法上的要点:当方程用于预测时,它们应该包含所有的变量,即使其中的一些系数可能并不显著地不等于 0,还是应该只包括那些具有显著地不等于 0 的系数的变量。

表 2.13 北爱尔兰社会剥夺程度的次序 logit 模型结果:新教徒样本

Ordered Logit Estimates Log Likelihood = −8124.0888; *Number of obs* = 8800;
LR $\chi^2(17)$ = 1527.75; *Prob*>χ^2 = 0.0000; *Pseudo* R^2 = 0.0859

y	*Coefficient*	*Standard Error*	*z*	*P*>|*z*|	[95% *Conf.*	*Interval*]
sex	−0.2293014	0.0444134	−5.163	0.000	−0.31635	−0.1422527
age	−0.0275073	0.0060613	−4.538	0.000	−0.0393873	−0.0156273
age2	0.0008768	0.0001131	7.754	0.000	0.0006551	0.0010984
ret	0.6275465	0.1008167	6.225	0.000	0.4299493	0.8251437
inac	1.179892	0.0954488	12.362	0.000	0.9928158	1.366968
ue	1.556987	0.0848854	18.342	0.000	1.390615	1.723359
highed	−0.8085732	0.085388	−9.469	0.000	−0.9759306	−0.6412159
mided	−0.6320822	0.1490549	−4.241	0.000	−0.9242244	−0.33994
hnum	0.95203	0.08401	11.332	0.000	0.7873735	1.116686
snpar	0.3647026	0.0845436	4.314	0.000	0.1990001	0.5304051
ard	−0.4509469	0.063093	−7.147	0.000	−0.5746069	−0.327287
crk	−0.3349415	0.0715481	−4.681	0.000	−0.4751733	−0.1947098
ant	0.3530984	0.0758826	4.653	0.000	0.2043712	0.5018256
col	0.368263	0.0833618	4.418	0.000	0.2048769	0.5316491
arm	0.4664409	0.0988147	4.720	0.000	0.2727676	0.6601141
ban	0.2097729	0.076176	2.754	0.006	0.0604708	0.359075
frm	0.6888737	0.0980013	7.029	0.000	0.4967947	0.8809528
_cut1	0.1488535	0.0834433			(Ancillary parameters)	
_cut2	2.04582	0.0869085				

y	*Probability*	*Observed*
1	Pr($xb+u<$ _cut1)	0.4970
2	Pr(_cut1 $<xb+u<$ _cut2)	0.3399
3	Pr(_cut2 $<xb+u$)	0.1631

表 2.14 北爱尔兰社会剥夺程度的次序 logit 模型结果:天主教徒样本

Ordered Logit Estimates Log Likelihood = −4283.363; *Number of obs* = 4364;
LR χ^2(17) = 875.06; *Prob*>χ^2 = 0.0000; *Pseudo* R^2 = 0.0927

y	*Coefficient*	*Standard Error*	*z*	*P* > |*z*|	[95% *Conf.*	*Interval*]
sex	−0.0430983	0.0621282	−0.694	0.488	−0.1648672	0.0786707
age	−0.0181642	0.0089546	−2.028	0.043	−0.0357148	−0.0006136
age2	0.00064	0.0001754	3.648	0.000	0.0002961	0.0009839
ret	0.6109521	0.1773859	3.444	0.001	0.2632821	0.958622
inac	1.487517	0.1299683	11.445	0.000	1.232784	1.742251
ue	1.364899	0.0906986	15.049	0.000	1.187133	1.542665
highed	−1.09291	0.1133639	−9.641	0.000	−1.315099	−0.8707208
mided	−0.7667852	0.2172056	−3.530	0.000	−1.1925	−0.3410701
hnum	0.9087931	0.0739094	12.296	0.000	0.7639333	1.053653
snpar	0.239649	0.0987062	2.428	0.015	0.0461884	0.4331096
ard	−0.5853662	0.16575	−3.532	0.000	−0.9102303	−0.2605022
crk	−0.3226664	0.1620424	−1.991	0.046	−0.6402637	−0.0050692
ant	0.2364382	0.1329838	1.778	0.075	−0.0242052	0.4970816
col	0.4795002	0.1039698	4.612	0.000	0.2757232	0.6832773
arm	0.2546794	0.0926459	2.749	0.006	0.0730968	0.4362621
ban	0.2025872	0.0954257	2.123	0.034	0.0155563	0.389618
frm	0.5243653	0.0950858	5.515	0.000	0.3380006	0.7107301
_cut1	0.0110645	0.1138485		(Ancillary parameters)		
_cut2	1.866844	0.1178967				

y	*Probability*	*Observed*
1	Pr($xb+u<$ _cut1)	0.3824
2	Pr(_cut1 $< xb+u <$ _cut2)	0.3682
3	Pr(_cut2 $< xb+u$)	0.2493

一种观点是,如果我们相信某个变量在方程设定中有一个合法的位置这一前提,那我们就应该坚持这个想法且不管怎样都把这个变量包括在内。但另一种观点是,因为估计与预测的目的是为了使方程设定面对数据,把预测基于从完整的模型设定中得到的系数估计值有可能具有误导性,因为它可能允许那些在模型设定中其合法性被数据明显“拒绝”的变量影响我们的预测。在前文中,追随第一种观点,预测基于完整的模型设定;在这一部分中,追随第二种观点,

预测基于受限制的模型设定。

表 2.15 说明:新教和天主教样本分别有 50%和 38%的人没有被剥夺,34%和 37%的人轻度被剥夺,16%和 25%的人严重被剥夺。与新教徒相比,天主教徒没有被剥夺的比例更小,轻度和严重被剥夺的比例更高(见表 2.15),这种实际情况可能有两个方面的原因:第一,那些使被剥夺的概率增加的特征不均衡地集中于天主教徒,以及/或者那些使被剥夺的概率下降的特征不均衡地集中于新教徒;[32]第二,某个特定的属性在具该属性的人为天主教徒而不是新教徒时,被更加严厉地惩罚(如果是能使剥夺增加的:比如处于失业状况),以及/或者不被更加慷慨地奖励(如果是能使剥夺降低的:比如具有教育资格)。

表 2.15 北爱尔兰社会剥夺程度:按宗教类别划分

	样本中的百分比		
	没有被剥夺	轻度被剥夺	严重被剥夺
总 体	45.9	34.9	19.2
天主教徒	38.2	36.8	24.9
新教徒	49.7	34.0	16.3

为了确定天主教—新教之间的剥夺差距(定义为在不同的剥夺程度上天主教徒与新教徒的比例差异)在多大程度上是由特征差异造成的,在多大程度上是由系数差异造成的,在下列问题上提出了计量经济学方面的议题:

1. 如果新教徒与天主教徒用各自的系数估计值评估他们具有的特征,那么对他们处于不同剥夺程度的预测概率会是怎样的?把这些概率分别记为 $\hat{p}_{ij}^{p}$ 和 $\hat{p}_{ij}^{c}$ ($i = 1, \cdots, 13164$; $j = 1, 2, 3$),它们的均值记为 $\bar{p}_j^p$ 和 $\bar{p}_j^c$ ($j = 1, 2, 3$)。[33]

2. 如果新教徒的特征用天主教徒的系数估计值进行评估,那么它们处于不同的剥夺程度的预测概率会是怎样的呢?把这些"合成的"概率记为 $\hat{q}_{ij}^{p}$ ($i = 1, \cdots, 13164$; $j = 1, 2, 3$),它们的均值记为 $\bar{q}_j^p$ ($j = 1, 2, 3$)。

3. $\bar{q}_j^p$ 与 $\bar{p}_j^c$ 合 $\bar{p}_j^p$ 相比怎样呢?

表 2.16 展示了作为一个整体时天主教徒与新教徒,以及在各自的次样本

中，这两组人的这三个概率 $\bar{p}_j^p$，$\bar{q}_j^p$ 与 $\bar{p}_j^c$ 的值。表 2.16 说明，对样本中所有个人进行计算，新教徒处于没有被剥夺，轻度被剥夺和严重被剥夺的平均“自己系数的”[34]概率分别为 50%，34%和 16%。当新教徒的特征用天主教徒的系数进行评估时，没有被剥夺的概率下降到 46%，严重被剥夺和轻度被剥夺的概率分别上升到 19% 和 35%。这个故事在其他被分析的次样本中重复(单身父/母亲、退休者、不从事经济活动者、失业者以及住在大家庭中的人)。当新教徒的特征用天主教徒的系数进行评估时，除了两个例外(退休者和失业者)，没有被剥夺的概率总是下降，严重被剥夺的概率总是上升，虽然很自然地，这些变化的大小随着考虑的次样本的不同而变化。[35](在没有被剥夺的概率中的)最大的下降与(在强烈被剥夺的概率中的)最明显的上升出现在那些不从事经济活动的人当中。

表 2.16 天主教徒和新教徒处于不同剥夺程度的概率估计值*

	概率估计值		
	没有被剥夺	轻度被剥夺	严重被剥夺
总体			
新教徒(用新教徒的系数估计)	49.7	34.0	16.2
新教徒(用天主教徒的系数估计)	45.8	35.3	18.9
天主教徒(用天主教徒的系数估计)	38.2	36.9	24.9
单身父/母亲			
新教徒(用新教徒的系数估计)	37.2	38.2	24.6
新教徒(用天主教徒的系数估计)	35.0	38.1	26.9
天主教徒(用天主教徒的系数估计)	27.3	37.8	34.8
退休者			
新教徒(用新教徒的系数估计)	24.8	41.4	33.8
新教徒(用天主教徒的系数估计)	24.6	40.7	34.6
天主教徒(用天主教徒的系数估计)	20.7	39.6	39.7
不从事经济活动者			
新教徒(用新教徒的系数估计)	31.3	41.4	27.3
新教徒(用天主教徒的系数估计)	20.7	38.9	40.5
天主教徒(用天主教徒的系数估计)	16.5	37.6	45.9
失业者			
新教徒(用新教徒的系数估计)	21.1	40.3	38.6

(续表)

	概率估计值		
	没有被剥夺	轻度被剥夺	严重被剥夺
新教徒(用天主教徒的系数估计)	21.3	39.5	39.2
天主教徒(用天主教徒的系数估计)	16.3	36.3	47.4
居住于大家庭者			
新教徒(用新教徒的系数估计)	28.1	40.0	31.9
新教徒(用天主教徒的系数估计)	26.0	39.0	34.9
天主教徒(用天主教徒的系数估计)	23.2	38.6	38.2

注:* 用表2.13和表2.14的次序logit估计结果计算而得。

在轻度被剥夺的概率方面的模式是,如果新教徒的特征用天主教徒的系数进行评估,在单身父/母亲,或退休者,或失业者,或居住在大家庭的人中,新教徒的位置不变,[36]但在那些不从事经济活动者中情况变差了。

新教徒和天主教徒在不同的剥夺类别中比例的差别,$\hat{p}_j^p - \hat{p}_j^c$ 可以分解为:

$$\bar{p}_j^p - \bar{p}_j^c = (\bar{q}_j^p - \bar{p}_j^c) + (\bar{p}_j^p - \bar{q}_j^p) = A_j^* + B_j^* \qquad [2.20]$$

在方程2.20中,A_j^* 表示在天主教徒与新教徒的"剥夺差异"中因这两组人之间特征的差别而导致的部分,B_j^* 表示由系数值的组间差别而导致的部分。这些绝对差异可以用比例表示:

$$A_j = A_j^* / (\bar{p}_j^p - \bar{p}_j^c) \text{ 以及 } B_j = 1 - A_j \qquad [2.21]$$

表2.17用百分比的形式,展示了 A_j 和 B_j 的值。

表2.17说明,对样本中所有个人进行计算,在那些没有被剥夺的天主教徒与新教徒之间,66%的剥夺差异($\bar{p}_j^p - \bar{p}_j^c$)是由于天主教次样本中的人具有与新教次样本中的人不相同的特征而引起的,34%是因为天主教徒特征的评估方式不同于新教徒特征的评估方式。至于严重被剥夺的情况,天主教徒—新教徒之间的差异有69%是由特征的不同造成的,但是在轻度被剥夺方面,只有55%的天主教徒—新教徒差异来源于特征差别。

转向具体的分组,对那些单身父/母亲、退休者、失业者来说,在不同的剥夺

程度上，相对来说较大比例的天主教徒—新教徒差异来自于两组人之间的特征差别。相对来说较小比例的差异是因为具体的一些特征，在应用到天主教徒身上时比应用到新教徒时，具有更加严重的剥夺后果。

表 2.17 对天主教徒和新教徒间剥夺程度差异的贡献率*

	百分比贡献率	
	特征差异(A_j)	系数差异(B_j)
总体		
没有被剥夺	65.8	34.2
轻度被剥夺	54.7	45.3
严重被剥夺	69.4	30.6
单身父/母亲		
没有被剥夺	78.2	21.8
轻度被剥夺	64.4	35.6
严重被剥夺	77.7	22.3
退休者		
没有被剥夺	96.1	3.9
轻度被剥夺	63.1	36.9
严重被剥夺	85.7	14.3
不从事经济活动者		
没有被剥夺	28.3	71.7
轻度被剥夺	33.9	66.1
严重被剥夺	29.4	70.6
失业者		
没有被剥夺	103.0	−3.0
轻度被剥夺	81.7	18.3
严重被剥夺	93.5	6.5
居住于大家庭者		
没有被剥夺	58.1	41.9
轻度被剥夺	27.2	72.8
严重被剥夺	51.4	48.6

注：* 用表 2.16 的概率估计值计算而得。

比如,对失业者来说,在严重被剥夺方面超过 94%的天主教徒—新教徒差异可以被这两组人之间特征的不同来解释。但是,对不从事经济活动者来说,被用于评估特征的系数的差别(相对于特征本身的差异)解释了超过 2/3 的天主教徒与新教徒之间的剥夺差异。对于住在大家庭中的人来说,在严重被剥夺方面几乎一半的组间差异,以及在轻度被剥夺方面超过 70%的组间差异,是由用于评估剥夺来源特征的系数的不同而引起的。

第 3 章 | 多类别模型

简介

前面考查了一组围绕着多元(>2)结果事件的模型，且这些结果在本质上是次序的。在这一系列模型里面，因变量 Y_i，在对个人 i 界定这些结果时(对第一种结果 $Y_i=1$；对第二种结果 $Y_i=2$；以此类推，直到对第 M 个(最后一个)结果 $Y_i=M$；$i=1, \cdots, N$)，是一个离散的、有序的变量。估计这些模型的合适的方法是次序 logit 和次序 probit。这一部分关注的是结果没有次序的多元结果模型。多类别 logit 的方法论——对这组模型的适合的估计方法——在对职业选择的分析中进行解释并加以应用。多类别 logit 的一个重要特征(和局限)是无关选项的独立性(IIA)。这一部分的结尾对这个特征以及怎样可能回避它的局限进行了讨论。

一些真实事件为我们提供了非次序结果的例子。工作交通工具(乘坐巴士、火车或者小汽车)的选择就是一个明显的例子。其他非次序结果的例子包括职业选择(非技术的、技术的、专业的、管理的)，居住地的选择(北边、南边、东边、西边)，以及选举时的政党选择(保守党、自由党、工党)。上述例子的一个共有的词语就是“选择”。因为这些结果并没有什么本质上的好或者坏，因此可以认为个人是从可得到的结果的选项中选择最合适他们的结果。其实，效用最大化的框架为理解非次序多元结果模型的结构提供了一个很好的切入点。

随机效用模型

给定 M 个选项(标记为 $j=1, \cdots, M$)中的一个选择，第 i 个人从第 j 个选

项中得出的效用可以表示为U_{ij}。假设这个效用是H个因素(决定变量)的一个线性函数。在这H个因素中,假设有R个是针对个人且与选择的属性无关的,S个($H=R+S$)是针对选择而与个人无关的。比如,在选择上班的交通方式方面,一个个人特征可能是他/她并不住在火车站附近。这个人几乎不坐火车去上班的事实与火车这个交通方式的属性没有关系,而仅仅是这个人住在哪里的结果。另一方面,高峰期交通非常繁忙这个实际情况是汽车出行的一个属性,它降低了所有人坐汽车上班的效用。假设代表了第i个人的特征的这R个变量的值为X_{ir}, $r=1, \cdots, R$,代表了第j种选择的属性的变量的值为W_{js}, $s=1, \cdots, S$。效用函数可以写成:

$$U_{ij} = \sum_{r=1}^{R}\beta_{jr}X_{ir} + \sum_{s=1}^{S}\gamma_{is}W_{js} + \varepsilon_{ij} = Z_{ij} + \varepsilon_{ij} \qquad [3.1]$$

其中β_{jr}是对第j种选择与第r个特征($r=1, \cdots, R$)相关的系数,γ_{is}是对第i个人与第s种属性($s=1, \cdots, S$)相关的系数;而且

$$Z_{ij} = \sum_{r=1}^{R}\beta_{jr}X_{ir} + \sum_{s=1}^{S}\gamma_{is}W_{js} \qquad [3.2]$$

个人i第r个特征的值X_{ir}的增加,会导致他/她从选择j中得到的效用上升,如果$\beta_{jr}>0$,下降,如果$\beta_{jr}<0$。选择j的第s种属性的值W_{js}的增加,对个人i会导致效用的上升,如果$\gamma_{is}>0$,下降,如果$\gamma_{is}<0$。但是,因为效用与它的决定变量之间的关系并不精确——比如,可能有些因素被排除在方程之外或者有些因素的测量不准确——所以方程中加入了一个误差项ε_i来捕捉这种不精确性。这个术语称为随机效用模型。

当而且只有当在所有可利用的选择中,第m种选择提供了最高程度的效用时,一个人才会选择$j=m$。也就是说,如果Y_i是一个随机变量,它的值$j=1, \cdots, M$表示个人i所做的选择,那么个人i会选择选项m的概率为:

$$\Pr(Y_i=m)=\Pr(U_{im}>U_{ij}) \quad \text{对所有 } j=1, \cdots, M, \text{且 } j\neq m$$

$$\Rightarrow \Pr(Z_{im}+\varepsilon_{im}>Z_{ij}+\varepsilon_{ij})$$

$$\Rightarrow \Pr(\varepsilon_{ij}-\varepsilon_{im}<Z_{im}-Z_{ij}) \quad \text{对所有 } j=1, \cdots, M, \text{且 } j\neq m$$

logit 模型的类别：多类别 logit 与条件 logit

麦克法登(McFadden，1973)已经证明过，如果 M 个误差项 ε_{ij} ($j = 1, \cdots, M$) 是以威布尔(Weibull)分布 $F(\varepsilon_{ij}) = \exp[\exp(-\varepsilon_{ij})]$ 独立同分布的，那么，

$$\Pr(Y_i = m) = \frac{\exp(Z_{im})}{\sum_{j=1}^{M}\exp(Z_{ij})} \qquad [3.3]$$

用方程 3.3 定义不同结果 $j = 1, \cdots, M$ 的概率的模型在这里称为广义 logit 模型。“广义”这个词表示模型同时包含了特征的作用和属性的作用，在方程 3.1 中分别为 X_{ir} 和 W_{js}。在广义 logit 模型这个类别里面，还可以再细分为两小类：

1. 多类别 logit 模型。这些模型只包含特征作用，所以对于它们来说方程 3.1 中所有的 $\gamma_{is} = 0$. 实际上，这些模型在针对个人的(individual specific)数据中使用。

2. 条件 logit 模型。这些模型只包含属性作用，所以对于它们来说方程 3.1 中所有的 $\beta_{jr} = 0$。实际上，这些模型在针对选项的(choice specific)数据中使用。

多类别 logit 模型

这个多类别模型由方程 3.3 定义但附带条件。现在，由于 $\gamma_{is} = 0$，

$$Z_{ij} = \sum_{r=1}^{R}\beta_{jr}X_{ir} \qquad [3.4]$$

因为 $\Pr(Y_i = j)$ 概率对所有的选择来说总和为 1(即 $\sum_{j=1}^{M}\Pr(Y_i = j) = 1$)，只有 $M-1$ 个概率可以独立决定。因此，方程 3.3 中的多类别 logit 模型是不定式，因为它是一个具有 M 个方程但只有 $M-1$ 个独立未知数的体系。解决这个问题的实用的标准化方法是设定 $\beta_{1r} = 0$, $r = 1, \cdots, R$。

在这个标准化下，$Z_{i1} = 0$，所以由方程 3.3：

$$\Pr(Y_i = 1) = \frac{1}{1 + \sum_{j=2}^{M} \exp(Z_{ij})} \qquad [3.5a]$$

$$\Pr(Y_i = m) = \frac{\exp(Z_{im})}{1 + \sum_{j=2}^{M} \exp(Z_{ij})} m = 2, \cdots, M \qquad [3.5b]$$

这个标准化的结果是，这些概率是唯一决定的，所以通过所采用的标准化方法，在方程 3.5a 已经确定了 $\Pr(Y_i = 1)$后，方程 3.5b 表示一个具有 $M-1$ 个方程和 $M-1$ 个独立未知数的概率 $\Pr(Y_i = m)$的体系。

从方程 3.5a 和 3.5b 中，结果 $j = k$ 与结果 $j = m$ 的概率比的对数为：

$$\log\left(\frac{\Pr(Y_i = m)}{\Pr(Y_i = k)}\right) = \sum_{r=1}^{R} (\beta_{mr} - \beta_{kr}) X_{ir} = Z_{im} - Z_{ik} \qquad [3.6]$$

那么这个风险比(risk-ratio)的对数(即结果 m 与结果 k 的概率比的对数，或者 $\log[\text{Prob}(Y_i = m)/\text{Prob}(Y_i = k)]$) 并不取决于其他的选择。这个"风险比"，或者有时候称为相对风险—— $\text{Prob}(Y_i = m)/\text{Prob}(Y_i = k)$ ——可以很容易地通过对风险比的对数取指数而计算出来。如果 $k = 1$，那么风险比的对数为：

$$\log\left(\frac{\Pr(Y_i = m)}{\Pr(Y_i = 1)}\right) = \sum_{r=1}^{R} \beta_{mr} X_{ir} = Z_{im} \quad (m = 2, \cdots, M) \qquad [3.7]$$

而风险比则为：

$$\frac{\Pr(Y_i = m)}{\Pr(Y_i = 1)} = \exp\left(\sum_{r=1}^{R} \beta_{mr} X_{ir}\right) = \exp(Z_{im}) \quad (m = 2, \cdots, M) \qquad [3.8]$$

要把这个风险比(RR)与比数比(OR)区别开来，后者指的是一个结果的概率除以 1 减去那个结果的概率。

因此，对 $j = m$，比率比为：

$$OR_m = \frac{\Pr(Y_i = m)}{1 - \Pr(Y_i = m)} = \frac{\Pr(Y_i = m)}{\Pr(Y_i = 1)} \frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = m)}$$

$$= \frac{RR_m \Pr(Y_i = 1)}{1 - RR_m \Pr(Y_i = 1)} \qquad [3.9]$$

其中，OR_m 和 RR_m 分别为与结果 $j = m$ 相关的比率比和风险比(后者以"基准"结果 $j = 1$ 为参照)。

在二分模型中，RR 与 OR 没有区别，因为基准结果 $Y_i = 1$ 就是 $Y_i \neq m$。在大于两种可能结果的模型中，结果 $Y_i = 1$ 与 $Y_i \neq m$ 是不一样的。所有 logit 模型的固有方法就是用一个结果与某个基准结果的可能性的比率表示分析结果——也就是，计算 RR。但是在二分模型中，RR 就是 OR，所以在这种模型中分析结果是用后者表示的。另一方面，在多类别 logit 模型中，由于 RR 与 OR 不一样，因此分析结果用 RR 表示。

估计与预测

因变量 $Y_i(i = 1, \cdots, N)N$ 种观测值中的每一个都视为从一个有 M 个结果的多项分布中的一次单一的抽取。定义虚拟变量 $\delta_{ij} = 1$，如果个人 i 选择 j，否则 $\delta_{ij} = 0$，$j = 1, \cdots, M$。则观察到这个样本的概率为：

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M}[\Pr(Y_i = j)]^{\delta_{ij}} \Rightarrow \log L = \sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{ij}\Pr(Y_i = j) \quad [3.10]$$

其中 $\Pr(Y_i = j)$ 如果 $j = 1$ 的话由方程 3.5a 定义，如果 $j > 1$ 的话由方程 3.5b 定义。参数估计值 $\hat{\beta}_{jr}(j = 1, \cdots, M, r = 1, \cdots, R)$ 的选择是为了使似然函数最大化（方程 3.10）。

给定这些估计值，对每个个人 $i = 1, \cdots, N$，我们可以使用方程 3.4，对每种结果 $j = 1, \cdots, M$ 用 $\hat{\beta}_{jr}$ 代替 β_{jr}，产生 Z_{ij} 的估计值。然后，使用这些估计值 $\hat{Z}_{ij}$，可以从方程 3.5a 和 3.5b 中计算出预测到的概率 $\hat{p}_{ij}$，$i = 1, \cdots, N$, $j = 1, \cdots, M$。注意对每个人来说，对所有的结果估计出来的概率必须总和为 1，$\sum_{j=1}^{M}\hat{p}_{ij} = 1$。多类别 logit 模型的一个特点就是，对每个结果估计出来的个人概率的平均数（$\bar{p}_j = \sum_{i=1}^{N}\hat{p}_{ij}$，$j = 1, \cdots, M$），等于实际观察到的在这个结果类别中的人所占的比例。

另一种从多类别 logit 模型中预测概率的方法是对每个结果 $j = 1, \cdots, M$，对所有人计算 Z_{ij} 的均值，如下：

$$\begin{aligned}\bar{Z}_j &= \sum_{i=1}^{N} Z_{ij}/N = \sum_{i=1}^{N}(\sum_{r=1}^{R}\beta_{jr}X_{ir})/N \\ &= \sum_{r=1}^{R}\beta_{jr}(\sum_{i=1}^{N}X_{ir}/N) = \sum_{r=1}^{R}\beta_{jr}\bar{X}_r \end{aligned} \quad [3.11]$$

然后计算估计到的概率如下：

$$\hat{p}_1 = \frac{1}{1+\sum_{j=2}^{M}\exp(\bar{Z}_j)} \qquad [3.12a]$$

$$\hat{p}_m = \frac{\exp(\bar{Z}_j)}{1+\sum_{j=2}^{M}\exp(\bar{Z}_j)} \quad m = 2, \cdots, M \qquad [3.12b]$$

通常，$\bar{p}_1 \neq \hat{p}_1$，$\bar{p}_2 \neq \hat{p}_2$，$\bar{p}_3 \neq \hat{p}_3$，这并不奇怪，因为虽然 $\bar{p}_j$ 和 $\hat{p}_j(j = 1, 2, 3)$两者都旨在测量不同结果的整体概率，但是它们的计算方法非常不同。$\bar{p}_j$ 计算为估计到的个人概率 $\hat{p}_{ij}$ 的均值，而 $\hat{p}_{ij}$ 是从个人决定变量的值中计算出来的。而另一方面，$\hat{p}_j$ 绕过了个人的概率，而用个人 Z_{ij} 的平均值 $\bar{Z}_j$ 计算(或者同等地，用决定变量 X_{ir} 的均值 $\bar{X}_r$ 计算)。

边际效应

决定变量的值的一个小变化对不同结果概率的边际效应这个问题可以用两种不同的方法表达。X_{ir} 的一个小变化(对个人 i 第 r 个决定变量的值)，对某个结果 m 在如下两点上会带来什么变化呢?

1. 概率 $\Pr(Y_i = m)$
2. 风险比 $\Pr(Y_i = m)/\Pr(Y_i = 1)$

第二个问题相对来说比较容易回答，但是第一个问题则难得多。先从简单的问题开始，由方程 3.7，

$$\frac{\partial}{\partial X_{ir}}\log\left(\frac{\Pr ob(Y_i = m)}{\Pr ob(Y_i = 1)}\right) = \beta_{mr}$$

因此对 X_{ir} 的一个小变化，风险比的变化方向可以从相关系数的符号中推断出来；如果 $\beta_{mr} > 0$ 则 $j = m$ 的相对概率增加，如果 $\beta_{mr} < 0$ 则减少。

但是，对 X_{ir} 的一个小变化，观察到结果 $j = m$ 的概率 $\Pr(Y_i = m)$其方向的变化，无法从 β_{mr} 的方向上推断出来。原因是，在多类别模型中，某个特定个人某变量值的变化会影响他/她每一种结果的概率。因为这些概率被限定总和为 1，$\Pr(Y_i = m)$上升还是下降取决于其他概率的变化情况。所以实际上它不

仅取决于 β_{mr} 的方向，还取决于该系数相对于与该变量相关的其他系数的大小情况，即相对于 $\beta_{jr} j = 1, \cdots, M, j \neq m$。因此，$\partial \Pr ob(Y_i = m) / \partial X_{ir}$ 不一定与 β_{mr} 的方向相同。

在多类别模型中，确定某变量值的变化对结果概率的影响的最有效的方法是，在其他变量的值保持不变的情况下，比较变化前与变化后的概率。当分析的决定变量是一个虚拟变量时，这种方法最有效。如果第 r 个变量是个虚拟变量，因此 $X_{ir} = 1$ 或者 $X_{ir} = 0$，那么首先假定对所有人 $X_{ir} = 1$，并用方程 3.4 对每个结果 $j = 1, \cdots, M$ 估计 Z_{ij}，把这个估计值称为 $\hat{Z}^1_{ij}$。然后，其他条件相同假定对所有人 $X_{ir} = 0$，对每个结果 $j = 1, \cdots, M$ 估计 Z_{ij}，把这个估计值称为 $\hat{Z}^0_{ij}$，注意 $\hat{Z}^1_{ij} = \hat{Z}^0_{ij} + \hat{\beta}_{jr}$。

用方程 3.5a 和 3.5b 计算概率的预测值，首先用 $Z_{ij} = \hat{Z}^1_{ij}$（计算出 $\hat{p}^1_{ij}$），然后用 $Z_{ij} = \hat{Z}^0_{ij}$（计算出 $\hat{p}^0_{ij}$）。对所有人计算平均概率，记为 $\bar{p}^1_j = \sum_{i=1}^{N} \hat{p}^1_{ij}$ 和 $\bar{p}^0_j = \sum_{i=1}^{N} \hat{p}^0_{ij}$。$\bar{p}^1_j$ 与 $\bar{p}^0_j$ 之间的差就是对所有人，第 r 个变量的值的变化对观察到结果 $j(j = 1, \cdots, M)$ 的概率的“平均”作用。

在分析某变量的值变化时，另外一种保持其他变量值不变的方法是把它们的值设为均值，$\bar{X}_r = \sum_{i=1}^{N} X_{ir} / N$，然后定义：

$$\bar{Z}^1_j = \beta_{jr} + \sum_{s \neq r} \beta_{js} \bar{X}_s \text{ 和 } \bar{Z}^0_j = \sum_{s \neq r} \beta_{js} \bar{X}_s \qquad [3.13]$$

然后用方程 3.5a 和 3.5b 计算概率的估计值，先用 $Z_{ij} = \bar{Z}^1_j$（计算出 $\hat{p}^1_j$），然后用 $Z_{ij} = \bar{Z}^0_j$（计算出 $\hat{p}^0_j$）。$\hat{p}^1_j$ 与 $\hat{p}^0_j$ 之间的差就是对“平均个体”（定义为对所有变量，其值都为平均值的个人），第 r 个变量的值的变化对这个人来说，观察到结果 $j(j = 1, \cdots, M)$ 的概率的作用。

某个决定变量其值的变化对风险比的影响要简单得多。从方程 3.7，$X_{ir} = 1$ 和 $X_{ir} = 0$ 的风险比对数的差别，为 $\hat{Z}^1_{ij} - \hat{Z}^0_{ij} = \hat{\beta}_{jr}$，所以对应的风险比的差别为 $\exp(\hat{\beta}_{jr})$。这样，某个系数的指数值表示的就是对某个决定变量一个单位的变化（对该结果）所带来的风险比的变化。当然，对虚拟变量来说，考虑一个单位的变化是最合适的。

应用：职业获得

方程设定

作者参考的是从英国1991普查中抽取出来的，关于25—45岁全职男性雇员的职业分类(以及其他特征)的信息。[37]职业的分类为非技术/半技术工人(UNS)、有技术体力/非体力劳动者(SKL)和专业/管理/技术人员(PMT)。

表3.1 男性全职雇员的样本统计值:按种族类别划分*

	白人	黑人	印第安人
样本量	96297	863	1572
年龄(年)	35.4	32.6	34.9
%职业类别			
专业/管理/技术人员(PMT)	41.7	29.3	35.1
有技术体力/非体力劳动者(SKL)	40.7	50.2	39.8
非技术/半技术工人(UNS)	17.6	20.5	25.2
%具18岁或以上的文凭			
学位文凭	15.1	7.9	17.4
副学位文凭	9.2	7.1	6.6
没有18岁或以上的文凭	75.7	85.0	76.0
%出生在			
英国	96.2	54.6	7.6
国外	3.8	45.4	92.4
%居住在			
伦敦	10.1	51.5	43.6
北部	50.3	24.9	30.2
南部	39.6	23.6	26.2

注：* 信息来自英国1991年人口普查。

在这98732个人当中，有96297个“白人”，863个“黑人”和1572个印第安人。表3.1显示了样本统计量的主要特征。这张表非常清楚地说明，职业地位在白人、黑人和印第安人之间存在着差别。比如说，42%的白种男性雇员的工

作属于 PMT 类别，但只有 35%的印第安人，以及 29%的黑人从事类似的工作。这些差别并不能被特征的差别解释掉：与 15%的白种工人相比，17%的印第安雇员有学位。

依据前面阐述的选择构成，看上去白人、黑人和印第安人在做选择时面临不同系列的限制。这可能是因为，与具有同等资格的白人来说，处于少数民族组别中的个人在劳动力市场上面临着劣势。也可能是因为与白人相比，处于少数民族组别中的个人具有更不利的劳工特征（此后称之为特征）。因此，职业表现的组间差别可能同时是种族与特征不利因素的结果。

关键问题是：这些差别中有多少是种族不利因素的结果，又有多少归因于特征的不利因素？前一部分所描述的多类别 logit 方法可以用来回答这个问题。

起点是对样本中每个 $i = 1, \cdots, 98732$ 男性定义因变量 Y_i 以使：

$Y_i = 1$ 如果这个人受雇于 UNS 职业类别；

$Y_i = 2$ 如果这个人受雇于 SKL 职业类别；

$Y_i = 3$ 如果这个人受雇于 PMT 职业类别。

在这个多类别模型设定中使用的决定变量包括：

年龄

AGE 以年计算：通过设定 25 岁的人 AGE = 0 进行标准化

教育[38]

HED = 1，如果这个人有学位程度文凭，否则 HED = 0；

MED = 1，如果这个人有高于 A 程度文凭（post-A level），但低于学位文凭；否则 MED = 0。

学习领域[39]

SCI = 1，如果为自然科学相关领域；否则 SCI = 0；

BUS = 1，如果为商学相关领域；否则 BUS = 0。

种族[40]

BLK = 1，如果这个人的种族为加勒比黑人，否则 BLK = 0；

IND = 1，如果这个人的种族为印第安人；否则 IND = 0。

出生国家

OVB = 1,如果这个人在英国以外出生;否则 OVB = 0。

居住地区[41]

STH = 1,如果住在英国南部[42](伦敦除外),否则 STH = 0;

NTH = 1,如果住在英国北部;[43]否则 NTH = 0。

以这些变量为基础,那两个风险比(对数)方程(方程 3.7,对 $j = 2, 3$)确定如下:

$$\begin{aligned}\log\left(\frac{\Pr(Y_i = j)}{\Pr(Y_i = 1)}\right) = {} & \alpha_{j0} + \beta_{j0} * BLK_i + \gamma_{j0} * IND_i + \alpha_{j1} * OVB_i \\ & + \beta_{j1} * BLK_i * OVB_i + \gamma_{j1} * IND_i * OVB_i \\ & + \alpha_{j2} * HED_i + \beta_{j2} * BLK_i * HED_i \\ & + \gamma_{j2} * IND_i * HED_i + \alpha_{j3} * MED_i \\ & + \beta_{j3} * BLK_i * MED_i + \gamma_{j3} * IND_i * MED_i \\ & + \alpha_{j4} * NTH_i + \beta_{j4} * BLK_i * NTH_i \\ & + \gamma_{j4} * IND_i * NTH_i + \alpha_{j5} * STH_i \\ & + \beta_{j5} * BLK_i * STH_i + \gamma_{j5} * IND_i * STH_i \\ & + \theta_{j1} * AGE_i + \theta_{j2} * AGE^2 + \theta_{j3} * BUS_i \\ & + \theta_{j4} * SCI_i + \theta_{j5} * BUS_i * HED_i \\ & + \theta_{j6} * SCI_i * HED_i \qquad [3.14]\end{aligned}$$

$j = 1$ 这个结果——也就是,处于 UNS 这个职业类别——在此后称为“基准”结果。这个结果的系数被设定为等于 0,其他结果的风险比定义为关于这个基准结果概率的风险比。[44]

在方程 3.14 中,α_{jr} 是“白人”的系数,β_{jr} 和 γ_{jr} 分别表示由于是黑人和印第安人,而对这些系数的附加的贡献。另一方面,θ_{jr} 是假定没有“种族维度”的系数,意即它们(被假定)在种族方面没有变化。从方程 3.14,对一个 25 岁(AGE = 0),英国出生的(OVB = 0),白种(BLK = IND = 0)男性雇员,且没有 18 岁或以上的文凭(HED = MED = 0),居住在伦敦(NTH = STH = 0),其风险比对数为

α_{j0}。如果有这么一个人是黑人或者印第安人，那么他的风险比对数会分别变化 β_{j0} 和 γ_{j0}，分别变为：$(\alpha_{j0}+\beta_{j0})$ 和 $(\alpha_{j0}+\gamma_{j0})$。

举例来说，如果 $\beta_{j0}<0$，那么处于职业类别 j 的概率相对处于 UNS 类别的概率的比率，对一个白人来说会比具有同等条件的黑人要高。系数 β_{j0} 和 γ_{j0} 的大小测量了黑人与印第安人相对于白人来说，(在职业类别 j 方面)面临的"种族不利因素"的程度，在如上描述的特征的情况下。[45] 方程 3.14 中的交互项，涉及种族变量 BLK 和 IND，允许种族不利因素的程度随着个人的某些非种族特征而变化。这些非种族特征为出生国家、居住地区和文凭程度。对具有上述特征的男性来说，处于职业类别 j 的风险比对在国外出生的(OVB = 1)白人来说是 $\alpha_{j0}+\alpha_{j1}$，对在国外出生的(OVB = 1)黑人来说是 $\alpha_{j0}+\alpha_{j1}+\beta_{j0}+\beta_{j1}$。

方程 3.14 包含的非种族特征交互项中，其中有一系列是学习领域和文凭程度之间的交互作用。这些交互项的相关系数为 θ_{j5} 和 θ_{j6}，且其值不等于 0。这意味着，在决定风险比的值时，不仅学习领域和文凭程度在分开考虑时有重要影响；而且学科与文凭是如何融合而产生作用也很重要，如一个具有自然科学相关学位的毕业生。

方程统计量

对方程 3.14 的估计，首先对它的系数没有附加任何限制，然后限定它的某些系数为 0。方程 3.14 一共包括 48 个系数，在 SKL($j=1$)和 PMT($j=2$)这两个方程中分别各占 24 个。在这 48 个系数中，有 13 个被设为 0。这些设为 0 的系数，各自并不显著地不等于 0，而且，似然比检验 $\chi^2(13)=8.1$，并没有拒绝它们全部等于 0 的联合假设。表 3.2 和表 3.3 分别表示，在没有附加任何限制和附加了零限制的情况下，对方程 3.14 的估计结果。对这些没有限制和有零限制的结果的比较显示，附加零限制并没有 S(quanlitatively affect)那些没有被设为 0 的系数的估计值。表 3.2 和表 3.3 中的 z-ratios 是估计系数相对于它们的估计标准误差的比率。在相关系数等于 0 的零假设下，这些 z-ratios 渐进地服从 $N(0, 1)$ 分布。[46]

表 3.2 职业选择的多类别 logit 估计结果：完整模型

Multinomial Regression Log Likelihood = −86948.078；*Number of obs* = 98732；
LR χ^2(46) = 30880.36；*Prob* > χ^2 = 0.0000；*Pseudo* R^2 = 0.1508

y	*Coefficient*	*Standard Error*	*z*	*P* > |*z*|	[95% *Conf. Interval*]	
y = 2						
age	0.0261899	0.0064884	4.036	0.000	0.0134729	0.0389069
age2	−0.0009772	0.000306	−3.194	0.001	−0.0015768	−0.0003775
blkcb	0.3655553	0.1672964	2.185	0.029	0.0376603	0.6934503
blkovs	0.08195	0.1887671	0.434	0.664	−0.2880266	0.4519266
ind	0.4707812	0.2790797	1.687	0.092	−0.0762051	1.017767
indovs	−0.6064709	0.2728072	−2.223	0.026	−1.141163	−0.0717786
ovsbn	−0.109104	0.0540097	−2.020	0.043	−0.214961	−0.003247
north	−0.1445068	0.0355612	−4.064	0.000	−0.2142055	−0.0748082
indnth	−0.4185296	0.1586727	−2.638	0.008	−0.7295225	−0.1075368
blknth	−0.7146398	0.2170033	−3.293	0.001	−1.139959	−0.2893211
south	−0.0776431	0.0365635	−2.124	0.034	−0.1493062	−0.0059799
indsth	−0.3459856	0.1654501	−2.091	0.037	−0.6702619	−0.0217094
blksth	−0.6664942	0.225652	−2.954	0.003	−1.108764	−0.2242244
highed	1.083631	0.1822909	5.945	0.000	0.7263477	1.440915
indhe	−0.2140724	0.4490211	−0.477	0.634	−1.094138	0.6659928
blkhe	−1.26384	0.7230326	−1.748	0.080	−2.680958	0.1532775
mided	0.6060102	0.2709862	2.236	0.025	0.0748871	1.137133
indme	−0.3332887	0.4428193	−0.753	0.452	−1.201199	0.5346213
blkme	0.9548953	1.034856	0.923	0.356	−1.073386	2.983177
subbus	1.205355	0.3189165	3.780	0.000	0.5802901	1.83042
subsci	0.5103959	0.2820694	1.809	0.070	−0.0424499	1.063242
bush	−0.6332213	0.4138681	−1.530	0.126	−1.444388	0.1779454
scih	−0.795496	0.3545109	−2.244	0.025	−1.490325	−0.1006674
_cons	0.7755227	0.042255	18.353	0.000	0.6927043	0.858341
y = 3						
age	0.0893431	0.0073103	12.222	0.000	0.0750152	0.103671
age2	−0.0024609	0.0003397	−7.244	0.000	−0.0031266	−0.0017951
blkcb	−0.0672642	0.1950314	−0.345	0.730	−0.4495187	0.3149904
blkovs	−0.5708287	0.2236755	−2.552	0.011	−1.009225	−0.1324329
ind	0.0918407	0.3348558	0.274	0.784	−0.5644646	0.748146
indovs	−0.7893394	0.3292557	−2.397	0.017	−1.434669	−0.1440101
ovsbn	0.0932756	0.0565007	1.651	0.099	−0.0174637	0.204015
north	−0.6113055	0.037637	−16.242	0.000	−0.6850727	−0.5375384
indnth	−0.4696822	0.1943706	−2.416	0.016	−0.8506416	−0.0887228
blknth	−0.1352359	0.2598729	−0.520	0.603	−0.6445774	0.3741057

（续表）

Multinomial Regression Log Likelihood = −86948.078; *Number of obs* = 98732; $LR\ \chi^2(46)$ = 30880.36; $Prob > \chi^2$ = 0.0000; *Pseudo* R^2 = 0.1508						
y	*Coefficient*	*Standard Error*	*z*	$P>\|z\|$	[95% *Conf. Interval*]	
south	−0.1697707	0.0383309	−4.429	0.000	−0.244898	−0.0946434
indsth	−0.5134623	0.1951871	−2.631	0.009	−0.8960221	−0.1309025
blksth	−0.418329	0.268549	−1.558	0.119	−0.9446753	0.1080174
highed	3.877128	0.1713991	22.620	0.000	3.541191	4.213064
indhe	−0.1698475	0.4062421	−0.418	0.676	−0.9660674	0.6263724
blkhe	−1.139974	0.6128429	−1.860	0.063	−2.341124	0.0611758
mided	2.850769	0.2490918	11.445	0.000	2.362558	3.33898
indme	−0.2049551	0.4154452	−0.493	0.622	−1.019213	0.6093026
blkme	0.743211	1.023254	0.726	0.468	−1.262329	2.748751
subbus	0.8151578	0.298235	2.733	0.006	0.230628	1.399688
subsci	0.2500103	0.2599529	0.962	0.336	−0.2594879	0.7595085
bush	−0.0414593	0.3898265	−0.106	0.915	−0.8055052	0.7225865
scih	0.1600076	0.3281777	0.488	0.626	−0.4832088	0.803224
_cons	−0.0164003	0.0462775	−0.354	0.723	−0.1071026	0.0743019

注：因变量 $y=1$ 为对照组。

表 3.3 职业选择的多类别 logit 估计结果：限制模型

Multinomial Regression Log Likelihood = −86952.126; *Number of obs* = 98732; $LR\ \chi^2(33)$ = 30872.27; $Prob > \chi^2$ = 0.0000; *Pseudo* R^2 = 0.1508						
y	*Coefficient*	*Standard Error*	*z*	$P>\|z\|$	[95% *Conf. Interval*]	
$y=2$						
age	0.026323	0.0064795	4.062	0.000	0.0136234	0.0390227
age2	−0.0009813	0.0003057	−3.210	0.001	−0.0015805	−0.0003822
blk	0.426683	0.1116093	3.823	0.000	0.2079328	0.6454332
blkovs	(dropped)					
ind	0.4144723	0.2152989	1.925	0.054	−0.0075059	0.8364504
indovs	−0.565894	0.2215624	−2.554	0.011	−1.000148	−0.1316396
ovsbn	−0.106519	0.05151	−2.068	0.039	−0.2074767	−0.0055613
north	−0.1438706	0.0351996	−4.087	0.000	−0.2128606	−0.0748806
indnth	−0.4152823	0.1546391	−2.685	0.007	−0.7183694	−0.1121952
blknth	−0.6568392	0.1761849	−3.728	0.000	−1.002155	−0.3115232
south	−0.0765746	0.0362333	−2.113	0.035	−0.1475906	−0.0055585
indsth	−0.3417737	0.1632409	−2.094	0.036	−0.66172	−0.0218275
blksth	−0.6677784	0.2086548	−3.200	0.001	−1.076734	−0.2588226
highed	1.373433	0.1088435	12.618	0.000	1.160104	1.586763

(续表)

Multinomial Regression Log Likelihood = −86952.126; Number of obs = 98732; LR χ^2(33) = 30872.27; Prob > χ^2 = 0.0000; Pseudo R^2 = 0.1508						
y	Coefficient	Standard Error	z	P > |z|	[95% Conf. Interval]	
indhe	(dropped)					
blkhe	−1.312484	0.7222905	−1.817	0.069	−2.728147	0.1031797
mided	0.8347366	0.1476385	5.654	0.000	0.5453704	1.124103
indme	(dropped)					
blkme	(dropped)					
subbus	0.9348005	0.1923779	4.859	0.000	0.5577467	1.311854
subsci	0.2707989	0.1355025	1.998	0.046	0.0052189	0.5363788
bush	−0.6003602	0.1675312	−3.584	0.000	−0.9287153	−0.272005
scih	−0.9608239	0.1601125	−6.001	0.000	−1.274639	−0.6470093
_cons	0.7737729	0.041826	18.500	0.000	0.6917954	0.8557504
y = 3						
age	0.0895118	0.0073026	12.258	0.000	0.0751989	0.1038247
age2	−0.0024658	0.0003395	−7.263	0.000	−0.0031313	−0.0018004
blk	(dropped)					
blkovs	−0.6781675	0.154667	−4.385	0.000	−0.9813092	−0.3750257
ind	(dropped)					
indovs	−0.7063556	0.1295174	−5.454	0.000	−0.9602051	−0.4525062
ovsbn	0.0954136	0.0553704	1.723	0.085	−0.0131104	0.2039377
north	−0.6105677	0.0368792	−16.556	0.000	−0.6828497	−0.5382858
indnth	−0.4542501	0.181704	−2.500	0.012	−0.8103834	−0.0981168
blknth	(dropped)					
south	−0.1682364	0.0376312	−4.471	0.000	−0.2419921	−0.0944807
indsth	−0.501347	0.1887684	−2.656	0.008	−0.8713262	−0.1313678
blksth	−0.4365967	0.2228379	−1.959	0.050	−0.8733511	0.0001576
highed	4.171942	0.0844821	49.383	0.000	4.00636	4.337524
indhe	(dropped)					
blkhe	−1.198389	0.6072452	−1.973	0.048	−2.388567	−0.0082099
mided	3.089609	0.0695418	44.428	0.000	2.95331	3.225909
indme	(dropped)					
blkme	(dropped)					
subbus	0.5326858	0.1343444	3.965	0.000	0.2693757	0.795996
subsci	(dropped)					
bush	(dropped)					
scih	(dropped)					
_cons	−0.0188423	0.0454545	−0.415	0.678	−0.1079314	0.0702469

注：因变量 $y=1$ 为对照组。

Mlogit：likelihood-ratio test $\chi^2(13)=8.1$；Prob>$\chi^2=0.8372$。

估计值

正如前面的讨论所强调的，表 3.2 和表 3.3 中系数估计值的方向反映了在其他条件同等的情况下，对应于该系数相关变量值的变化，风险比 $\Pr(Y_i = j)/\Pr(Y_i = 1)$ 的变化方向。它并不反映单个概率 $\Pr(Y_i = j)$ 的变化方向。在这一部分中报告的估计结果是表 3.3 中的约束系数。这是因为，当把完整的方程设定(见表 3.2)应用到数据中时，我们发现这些变量只有一小部分对风险比有显著的影响。在大多数情况下，被排除在外的变量是那些涉及某个非种族变量 X 与种族变量 BLK 和/或 IND 的交互项。这意味着当变量 X 本身对个人 i 属于某个特定职业类别的风险比有显著影响时，这个人属于哪个种族并不会改变这种影响。

表 3.3 所示的估计结果，点出了三种对改善个人处于 SKL 或 PMT 职业类别风险比来说很重要的特征：[47]

住在伦敦；

具有 18 岁或 18 岁以上的文凭，如果取得学位或学习与商业相关的科目会更好；

出生在英国。

住在伦敦有两点好处：首先，对所有人有一个整体的益处。这来源于如下情况：居住在北部或者南部的男性，在其他条件相同时，与居住在伦敦的人相比，其处于 SKL 或 PMT 职业类别的风险比更低[48](方程 3.14 中 $\hat{\alpha}_{j4}$，$\hat{\alpha}_{j5} < 0$，$j = 2, 3$)。其次，对黑人和印第安人来说还有特定的好处。因为住在伦敦而不是在伦敦之外，黑人和印第安人与白人相比，在处于 SKL 或 PMT 职业类别的风险比上具有更大的提升。相比于住在伦敦，住在伦敦之外的人其处于 SKL 或 PMT 职业类别的风险比的下降程度，对黑人[49]或者印第安人比对白人更大($\hat{\beta}_{j4}$，$\hat{\beta}_{j5}$，$\hat{\gamma}_{j4}$，$\hat{\gamma}_{j5} < 0$，$j=2$ 以及 $\hat{\beta}_{j5}$，$\hat{\gamma}_{j4}$，$\hat{\gamma}_{j5} < 0$，$j = 3$)。

从这种意义上说，相对于白人，伦敦比英国其他地区对印第安人和黑人更加友善。特别是住在伦敦的印第安人($NTH_i = STH_i = 0$)，并不比白人在处于 PMT 职业类别的风险比方面具有劣势，因为 $\hat{\gamma}_{j0} = 0$，$j = 3$。住在伦敦之外($NTH_i = 1$)对所有人这个风险比都下降，但是对印第安人来说下降程度要比

对白人更大。从这种意义上说，在 PMT 职业类别中，印第安人在伦敦相对于白人所享受的种族同等情况，在伦敦以外的地区减弱了。“伦敦效应”对黑人最有利。在全职男性雇员中，只有 10%的白人住在伦敦，但有 44%的印第安人和 53%的黑人住在这个地区。

出生在英国之外一直是一个劣势：处于 SKL 或 PMT 职业类别的风险比对英国出生的人比在国外出生的人更大（方程 3.14 中 $\hat{\alpha}_{j1} < 0$，$j = 2, 3$）。但是，因为出生地这个变量与种族的交互项是负的（$\hat{\beta}_{j1}\hat{\gamma}_{j1} < 0$，$j = 2, 3$），出生在国外的劣势对印第安人和黑人来说，要比白人更大，即由于出生在国外而导致的风险比的降低程度对印第安人和黑人比白人更大。

拥有 18 岁或 18 岁以上的文凭，不管是副学位还是学位（或更高）文凭，两者都提高了处于 SKL 或 PMT 职业的风险比，其中学位文凭的作用比副学位文凭的作用更大。

除了文凭之外，所取得文凭的科目也有重要作用。商业类科目的文凭为进入 SKL 和 PMT 这两种职业都提供了最好的途径。计量经济学的结果是对男性印第安人来说，方程 3.14 中的 γ_{j2}（HED）和 γ_{j3}（MED）在 SKL（$j = 1$）和 PMT（$j = 2$）这两种职业中都不显著地不等于 0。对男性黑人来说，在 PMT 职业类别中 β_{j2}（HED）明显是负的。男性黑人相比印第安人或白人来说，具有更低的学位文凭回报率，但是印第安人和白人的回报率相同。

相关文献指出，那些在自己的国家获得文凭的移民，尤其是从发展中国家来的移民，持有这样一种观念（不管是否有理）：这些文凭与在接收国家获得的同等文凭相比，更“没有价值”。遗憾的是，我们的数据并没有对那些出生在国外的人，记录他们迁入英国的时间。因此，虽然我们知道样本中大部分黑人和印第安人出生在国外（表 3.1），但是却没有关于他们到达英国时的年龄方面的信息，这意味着，我们无法推断他们在哪里获得他们的 18 岁或 18 岁以上的文凭（如果有的话）。

概率估计值

利用估计出来的 $\hat{Z}_{ij}$——记住 $\hat{Z}_{ij} = \sum_{r=1}^{R} \hat{\beta}_{jr} \times X_{ir}$ 是用表 3.3 中的估

计值和每个人的决定变量的值联合估计出来的——STATA 会通过计算方程 3.5a中的 $\hat{p}_{i1}$和 3.5b 中的 $\hat{p}_{i2}$和 $\hat{p}_{i3}$，对样本 98732 个人中的每个人估计处于三种不同职业类别的概率。对所有个人，以及对白人、黑人和印第安人这三组人，分别计算出来的这些个人概率的均值，分别记为 $\bar{p}_j$，$\bar{p}_{Wj}$，$\bar{p}_{Bj}$和 $\bar{p}_{Ij}$ (j = 1, 2, 3)。如表 3.4 所示，这些平均概率(估计到的个人概率的均值)表示，比如，对所有男性雇员和白人，黑人以及印第安男性雇员估计到的处于 PMT 职业类别的平均概率分别为 41.5%，41.8%，29.3%和 35.1%。应该强调的是，对所有人和对每个种族组别内的个人所估计到的处于三种职业类别的平均概率，正好等于所对应的每个分类的样本比例。[50]

表 3.4　白人、黑人和印第安人处于不同职业类别的概率估计值*

	概率估计值		
	UNS	*SKL*	*PMT*
个人概率估计值的均值			
总体	17.7	40.8	41.5
白人	17.5	40.7	41.8
黑人	20.5	50.2	29.3
印第安人	25.1	39.8	35.1
用决定变量的均值计算而得的概率估计值			
总体	14.3	41.8	43.9
白人	14.2	41.7	44.1
黑人	19.6	52.0	28.4
印第安人	22.0	43.3	34.7

注：* 用表 3.3 的多类别 logit 估计结果计算而得。
UNS = 非技术/半技术工人。
SKL = 有技术体力/非体力劳动者。
PMT = 专业/管理/技术人员。

这是多类别 logit 模型的一个特点，估计到的个人结果概率的均值总是等于这些结果的样本比例。

估计概率的另一种方法是对所有人，以及对白人、黑人和印第安人这三组人计算$\hat{Z}_{ij}$的均值。如果这些均值分别记为$\bar{Z}_j$，$\bar{Z}_{Wj}$，$\bar{Z}_{Bj}$，$\bar{Z}_{Ij}$，那么方程3.12a

就可以用来对所有人和每一组人计算第一种结果的概率 $\hat{P}_1$，$\hat{P}_{W1}$，$\hat{P}_{B1}$ 和 $\hat{P}_{I1}$)，方程 3.12b 就可以用来对所有人和每一组人计算第二种和第三种结果的概率 $\hat{P}_j$，$\hat{P}_{Wj}$，$\hat{P}_{Bj}$ 和 $\hat{P}_{Ij}$ ($j = 2, 3$)。[51]这些概率，如表 3.4 所示(用决定变量的均值估计的概率)，显示在决定变量为均值时，所有人、白人、黑人和印第安男性雇员处于 PMT 职业的概率估计值分别为 43.9%，44.1%，28.4%和 34.7%。

这些概率估计值与该表上半部分报告的概率不一样，这并不奇怪。举例来说，虽然 $\bar{p}_{Wj}$ 和 $\hat{P}_{Wj}$ ($j = 1, 2, 3$)两者都旨在测量白人处于不同职业类别的总体概率，但是他们的计算方法非常不同。$\bar{p}_{Wj}$ 是(对 96297 个白人)估计到的个人概率 $\hat{p}_{ij}$ 的均值，而 $\hat{p}_{ij}$ 是用个人决定变量的值计算出来的。另一方面，$\hat{P}_{Wj}$——绕过了个人概率——反而是用样本 96297 个白人 Z_{ij} 的均值 $\bar{Z}_{Wj}$ 计算出来的。[52]因此，总的来说，对任何组别 g，$\bar{P}_{g1} \neq \hat{P}_{g1}$；$\bar{P}_{g2} \neq \hat{P}_{g2}$；$\bar{P}_{g3} \neq \hat{P}_{g3}$。

通过模拟获取边际效应

在之前的讨论中，我们强调了对 X_{ik} 的一个小变化观察到结果 j 的概率 $\Pr(Y_i = j)$ 的变化方向不能从 β_{jr} 的方向中推断出来，因为$\partial \text{Prob}(Y_i = j)/\partial X_{ir}$不一定与 β_{jr} 的方向相同。在多变量 logit 模型的情况下，只有风险比的变化方向能从系数的方向中估计出来。因此，前面对估计结果的讨论就是关于这些风险比的。

但是，可能有人会对概率 $\Pr(Y_i = j)$而不是相对于某个基准结果的风险比 $\Pr(Y_i = j)/\Pr(Y_i = 1)$感兴趣。在这个应用实例中，可能有人会对处于某个具体职业类别的概率在不同的种族间如何存在差别特别感兴趣。由于系数估计值并不直接对此给出答案，一种可以替代的方法就是通过模拟的窗口观察这些结果。我们可以通过计算在不同的假设情形下不同结果的概率而使模型的结果变得清晰明显。[53]更具体地，种族的作用可以通过比较当种族虚拟变量取不同值时的概率分析出来，在做这些比较时其他变量的值保持不变。

假设样本中所有 98732 个人都是白人，即对所有 $i = 1, \cdots, N$，$IND_i = BLK_i = 0$。让 P_{ij}^W 表示在这个假设情形下个人 i 处于职业类别 j ($j = 1, 2, 3$)

的概率。现在假设样本中所有的 98732 个人都是黑人，即对所有 $i=1, \cdots, N$, $BLK_i=1$，让 P_{ij}^B 表示在这个假设情形下个人 i 处于职业类别 j ($j=1, 2, 3$)的概率。[54]根据方程 3.14 的设定，这两个概率 P_{ij}^W 和 P_{ij}^B 是当样本中所有人的特征分别用白人(α_{jr})和黑人的系数($\alpha_{jr}+\beta_{jr}$)进行评估时的结果。此后称它们为"种族"概率，因为它们是用只在种族特征上不同的个人所组成的样本计算出来的。当应用到方程 3.5a 和 3.5b 时，系数估计值和 $\hat{\alpha}_{jr}$ 和 $\hat{\alpha}_{jr}+\hat{\beta}_{jr}$ 产生种族概率的估计值 $\hat{P}_{ij}^W$ 和 $\hat{P}_{ij}^B$。

如果 $\bar{p}_j^W=\sum_{i=1}^N \hat{P}_{ij}^W/N$, $\bar{p}_j^B=\sum_{i=1}^N \hat{P}_{ij}^B/N$ 是个人概率估计值各自的均值(其中后者是在两组假定的情形下计算出来的)。那么，种族概率的均值 $\bar{p}_j^W$ 和 $\bar{p}_j^B$ 之差就完全是应用于(样本中 N 个人)给定特征时，不同系列的系数(白人和黑人)之间差别的结果。因此，这些差异可以完全归因于对这两组人不平等对待的结果，而这两组人除了种族不同，在其他每个方面都完全相同。更简明地说，可以归因于黑人遭受"种族劣势"这个事实。类似的应用也可以在假定样本中每个人都为印第安人的情况下实行。

这些种族概率的估计值，分别对白人、黑人和印第安人这三组人，见表 3.5 的上面那部分。这部分显示如果样本中每个人的特征都用黑人的系数 $\alpha_{jr}+\beta_{jr}$，而不是用与他们相关的种族系数估计，[55]那么总样本中 40%的人会在 PMT 职业类别中，39%会在 SKL，21%会在 UNS 职业中。

另外一种用这个方式计算种族概率的方法是，比较当种族变量取不同的值但其他变量都在各自情形下取样本均值时的概率。用 $\overline{Z}_j^W$ 和 $\overline{Z}_j^B$ 表示当样本中所有人都(被视为)是白人和样本中所有人都(被视为)是黑人时 $\hat{Z}_{ij}=\sum_{r=1}^R \hat{\beta}_{jr}X_{ir}$ 的均值。[56]

那么白人和黑人处于不同职业类别中的概率(分别用 $\hat{P}_j^W$ 和 $\hat{P}_j^B$, $j=1, 2, 3$ 表示)是用 $\overline{Z}_j^W$ 和 $\overline{Z}_j^B$ 计算出来的，对第一种职业类别用方程 3.12a，对第二和第三种用方程 3.12b。这些估计出来的概率的值，分别对白人、黑人和印第安人列于表 3.5。当概率是用样本的平均特征计算时，估计到更大比例的白人和黑人(更小比例的印第安人)处于 PMT 类别($\hat{P}_3^W=44.2\%$, $\hat{P}_3^B=41.0\%$,

$\hat{P}_3^I = 32.6\%$)，相对于当概率是用个人概率的均值计算时($\bar{P}_3^W = 41.8\%$，$\bar{P}_3^B = 39.8\%$，$\bar{P}_3^I = 33.7\%$)。这是因为印第安人具有比均值更高的高等教育文凭特征(见表3.1)，而这种特征对获得PMT职业来说是成功的主要决定因素。因此所有人都降低到样本均值时，印第安人会“损失”而其他人则“获利”。

但是，正如表3.5(样本比例)所示，只有29%的黑人男性雇员实际上在PMT职业内。黑人和白人在职业类别j的样本比例，分别用s_j^B和s_j^W表示，通常不同于种族概率，$\bar{P}_j^B$和$\bar{P}_j^W$(以及$\hat{P}_j^B$和$\hat{P}_j^W$)。这反映了一个情况，即黑人和白人不仅在他们是如何被对待的这方面不同，而且在他们的特征方面也不同。黑人处于PMT职业的样本比例($s_3^B = 29.3\%$)比估计到的处于该职业的种族概率($\bar{P}_3^B = 39.8\%$或者$\hat{P}_3^B = 41.0\%$)低的情况是因为，相对于整体样本，黑人的特征更不适合PMT职业这个实际情况。或者说，黑人相对于白人，在PMT职业方面会同时遇到种族和特征方面的劣势。这些单独的劣势之和被称为总体劣势。

表3.5 白人、黑人和印第安人处于不同职业类别的“种族”概率估计值*

	概率估计值		
	UNS	*SKL*	*PMT*
个人概率估计值的均值			
白人	17.5	40.7	41.8
黑人	20.7	39.5	39.8
印第安人	19.3	47.0	33.7
样本比例			
白人	17.5	40.7	41.8
黑人	20.5	50.2	29.3
印第安人	25.1	39.8	35.1
用决定变量的均值计算而得的概率估计值			
白人	14.2	41.6	44.2
黑人	19.3	39.7	41.0
印第安人	16.6	50.8	32.6

注：* 用表3.3的多类别logit估计结果计算而得。
UNS = 非技术/半技术工人。
SKL = 有技术体力/非体力劳动者。
PMT = 专业/管理/技术人员。

测量职业劣势

黑人同白人相比，在职业类别 j 中遇到的平均的种族劣势的一个测量指标是 λ_j^B，其中：

$$\lambda_j^B = \bar{p}_j^B / \bar{p}_j^W \quad [3.15]$$

如果这两个概率相等，则 $\lambda_j^B = 1$，没有种族劣势。但是，如果 $\bar{p}_j^B < \bar{p}_j^W$，则 $\lambda_j^B < 1$，对职业类别 j 存在黑人种族劣势[57]，λ_j^B 离 1 越远，则这种劣势的规模越大。

黑人同白人相比，在职业类别 j 中遇到的平均的总体劣势的一个测量指标是 μ_j^B，其中：

$$\mu_j^B = s_j^B / s_j^W \quad [3.16]$$

如果这两个样本比例相等，则 $\mu_j^B = 1$，没有总体劣势。但是如果 $s_j^B < s_j^W$，则 $\mu_j^B < 1$，对职业类别 j 存在黑人总体劣势[58]，μ_j^B 离 1 越远，则这种劣势的规模越大。

黑人同白人相比，在职业类别 j 中遇到的平均特征劣势的一个测量指标是 δ_j^B，其中 δ_j^B 是总体劣势与种族劣势的比：$\delta_j^B = \mu_j^B / \lambda_j^B$。

如果 $\delta_j^B = 1$，则黑人没有面临特征劣势，因为 $\mu_j^B = \lambda_j^B$。这种情况下，黑人与白人样本比例的比等于对应的种族概率的比。如果后面这个比小于 1，则完全归因于黑人与白人相同的特征被区别对待，对后者的评估比对前者更加有利。但是，如果 $\delta_j^B < 1$，则存在特征劣势，而且 δ_j^B 离 1 越远这种劣势的规模越大。黑人在他们的特征方面受到惩罚 $\mu_j^B < \lambda_j^B$。他们因为具有较差的特征而吃亏，此外这些特征还(可能)更加不利地被评估，相比假若它们是白人的特征的话。最后，如果 $\delta_j^B > 1$，则黑人享受了特征优势 $\mu_j^B > \lambda_j^B$。黑人因为具有更好的特征而弥补了可能的不公平对待所带来的损失。其实，如果 δ_j^B 的值足够大那么就可能出现 $\mu_j > 1$，即使 $\lambda_j < 1$。换句话说，黑人较有优势的特征可以甚于中和他们可能遇到的任何不公平对待。表 3.6 展示了三个不同种族的这三种不同

的劣势估计值。

表 3.6 黑人和印第安人相对于白人面临的种族、特征和总体劣* 的估计结果

	种族劣势(%)	特征劣势(%)	总体劣势(%)
用个人概率估计值的均值计算而得			
黑人/白人			
$j=1$(SKL)	0.97	1.27	1.23
$j=2$(PMT)	0.95	0.74	0.70
印第安人/白人			
$j=1$(SKL)	1.15	0.85	0.98
$j=2$(PMT)	0.81	1.04	0.84
用决定变量均值计算而得的概率估计值计算而得			
黑人/白人			
$j=1$(SKL)	0.95	1.26	1.23
$j=2$(PMT)	0.93	0.75	0.70
印第安人/白人			
$j=1$(SKL)	1.22	0.80	0.98
$j=2$(PMT)	0.73	1.15	0.84

注：* 若取值 >1，则为具有优势。用表 3.5 中的数字计算而得。
UNS = 非技术/半技术工人。
SKL = 有技术体力/非体力劳动者。
PMT = 专业/管理/技术人员。

表 3.6 中“种族劣势”这一栏下面的数字表示，如果印第安人和白人配以一组共同的特征(即总样本的特征)的话，那么印第安人处于 PMT 类别的概率等于对应白人概率的 81%。黑人的类似经历会产生 95%这个数值。但实际上，印第安人和黑人作为不同的组别，具有与总体样本不同的特征，这意味着印第安人和黑人在 PMT 类别的样本比例，分别为对应白人比例的 84%和 70%。印第安人从种族的 81%这个数值上升到样本的 84%这个数，以及黑人对应地从 95%下降到 70%，可能分别归因于，印第安人“具有优势的”特征和黑人“具有劣势的”特征。表 3.6 显示，在 PMT 职业方面，相对于白人，印第安人有 4%的特

征优势，但是黑人则有 15%的特征劣势。

在 SKL 这个类别，印第安人相对于白人具有 15%的特征劣势，而黑人相对于白人则有 27%的特征优势。

我们也可以把种族劣势定义为 $\hat{P}_j^B/\hat{P}_j^W$——即通过设定除了种族之外的其他变量等于它们的样本均值而获得的概率比(见表 3.5 的下部分)。特征劣势现在重新计算为总体劣势[59]与这种新的种族劣势估计值之间的比。新的种族和特征劣势的值(以及没有变化的总体劣势的值)见表 3.6。

条件 logit 模型与不相关选项的独立性

在条件 logit 模型中，结果概率 $\Pr(Y_i = j)$ 仅仅取决于选择的属性而不取决于做选择的个人的特征。或者说，条件 logit 模型由方程 3.3 定义但附带条件 $\beta_{jr} = 0$，

$$Z_{ij} = \sum_{s=1}^{S} \gamma_{is} W_{js} \qquad [3.17]$$

除此之外，这个模型与多类别 logit 模型完全一样。通常忽略个人特征的原因是没有个人数据。因此，条件 logit 模型可以有效地看做个人的集合体——此后称做“总体”——在具有不同属性的选项之间做选择——比如，上下班的人选择不同的交通方式或者消费者选择不同的超级市场。通过这种想法，与个人相关的下标 i 就可以在下面的讲解中去掉了。因变量 $Y = j$ 与总体所做的选择相关，系数 γ_s 指总体附于第 s 个属性的权重。与在多类别模型中一样，在条件 logit 模型中任意两个选项 j 和 k 的风险比独立于其他可供选择的选项，

$$\log\left(\frac{\Pr(Y_i = j)}{\Pr(Y_i = k)}\right) = Z_j - Z_k = \sum_{s=1}^{S} (W_{js} - W_{ks})\gamma_s \qquad [3.18]$$

在方程 3.18 中，风险比对数仅仅取决于与 j 和 k 相关的属性(间的差别)，与其他选项的属性无关。[60]这个特性称为不相关选项的独立性(IIA)，既是这些模型的主要优点，也是它们的主要缺点。[61]它是优点因为它允许在不需要重新估计模型的情况下引进新的选项。

比如,假设引进了第$(M+1)$个选项,其属性为$W_{M+1,s}s=1, \cdots, S$。与这个新选项相关的估计到的 Z 值(Z-value)为:$\hat{Z}_{M+1}=\sum_{s=1}^{S}\hat{\gamma}_s W_{M+1,s}$,其中系数估计值 $\hat{\gamma}_s$ 已经有了。新选项的采用所需要的只是,在方程 3.3 的分子中加入一个新的项 $\hat{Z}_{M+1}$并对概率重新计算。已有的系数估计值能同时在选项扩充之后使用的原因在于,新选项的加入并不能改变已有的选项被选择时的相对风险。

假定加入了某个新选项,选择特定选项的概率其变化的百分比可以很容易计算为:

$$\begin{aligned}\frac{\Pr(Y^{M+1}=j)-\Pr(Y^{M}=j)}{\Pr(Y^{M}=j)} &= \left(\frac{\exp(Z_j)}{\sum_{j=1}^{M+1}\exp(Z_j)}\frac{\sum_{j=1}^{M}\exp(Z_j)}{\exp(Z_j)}\right)-1\\ &= \frac{-\exp(Z_{M+1})}{\sum_{j=1}^{M+1}\exp(Z_j)}\\ &= -\Pr(Y^{M+1}=M+1)\end{aligned}$$

其中,$\Pr(Y^{M+1}=j)$和$\Pr(Y^{M}=j)$分别指当有$M+1$个和M个可供选择的选项时,选择选项 j 的概率。因此引进一个额外的选项时,选择已有的不同选项的概率将会减少相同的百分比,这个百分比等于选择这个新选项的概率。

但是,这个在引进新选项时已有概率下降相同百分比的特性同时也是一个缺点,因为它意味着对已有选项,关于新选项的需求的交叉弹性在不同的选项间是一致的。[62]要使这个有效,这些选项必须被看做完全不同且独立的。它们被如此看待并导致 IIA 这个实际情况,源于误差是独立同分布的这个假设。

该模型的局限性有多大的负面影响取决于所分析问题的本质。比如,在职业选择的模型中,选项可能被视为定义明确且相对不可改变,引进新选项带来的问题可能并不严重。[63]但另一方面,在选择上下班交通方式这个例子中,必须考虑到新的交通方式的引进——新的火车路线、公共汽车路线、小车交通拥挤费,等等——极有可能带来严重的问题。

一个典型的由 IIA 假设引起的问题的例子是“红巴士—蓝巴士”问题。假设

上下班者有三种选择:乘坐小汽车、火车或者巴士去上班。顺便地假设所有的巴士都漆成红色,并且 $\Pr(Y=j)$ 的 logit 估计值为:

小汽车:55.4

巴士(红色):23.0

火车:21.6

小汽车和巴士这两个交通方式间的风险比为 $55.4/23.0=2.4$。现在巴士公司决定把一半的车漆成蓝色,我们可以很合理地预期这个纯粹表面的改变并不会给通勤者的选择带来任何变化,并且相同百分比的人,如上所示,将会继续使用小汽车、巴士和火车上班。但是,由于模型不能识别虚假的和真实的新选项,在引进"第四种"选项后预测到的 logit 概率变成[64]:

小汽车:45.1

巴士(红色):18.8

巴士(蓝色):18.4

火车:17.7

注意,每个已有选项小汽车、(红色)巴士和火车的概率都下降了 18.4%,即使用蓝巴士上班的概率。使用小汽车和红色巴士(以及使用火车和红色巴士)之间的风险比保持不变,但是为了迁就这个情况,一些之前使用小汽车、火车或者红色巴士的通勤者必须改变以乘坐蓝色巴士上班。因此,仅仅由于把一半的车改漆成蓝色,巴士公司就可以使所有通勤者中乘坐巴士上班的比例由 23%增加到 37.2%。

另一个更重大的例子可以从政界中找到。假设某个政党 X 与其他两个政党 Y 和 Z 竞争。用多类别 logit 模型我们可以用相对于投 X 政党票的概率的一个比值,来估计支持 Z 的投票者的概率。假设现在政党 X 分成两个党,$X1$ 和 $X2$,它们具有类似的(比如左派的)观点,且笼统地被投票者认为两者间并没有本质差别。但是,由于模型不能识别虚假的和真实的新选项,在引进"第四种"政党后,预测到的投左派政党的票的 logit 概率将会大幅上升。这就是条件 logit 的一大缺陷:在 IIA 假设下,模型无法解决由引进类似的或相同的选项所带来的虚假的比例膨胀问题。

顺便提一下,我们应该注意由混合选择所带来的更进一步的局限性——比如,对上班时间和交通模式两者同时进行的选择。那么,正如多门西克和麦克法登(Domencich & McFadden, 1996)指出的,选择时间 t 和模式 j 的联合概率 $\Pr(t \cap j)$,可以写成(在时间 t,从可供利用的选择模式中,选择模式 j 的概率)和(从可供利用的时间选项中选择时间 t 的概率)的联合结果。换句话说,IIA 的一个含义是,包含混合选择的效用函数必须是可以从个人选择中加法可分的,$U(t, j) = \phi(j) + \psi(t)$。

logit 模型的替代模型

在考虑 logit 模型的替代方法之前,重要的是先检查 IIA 假设是否有效。假设我们相信选择的某个子集是不相关的。把它从选择集中去掉应该不会显著改变系数估计值。这个想法是豪斯曼和麦克法登(Hausman & McFadden, 1984)检验的依据。如果 $\hat{\gamma}_1$ 是基于受了限制的选择项而得到的系数估计值的向量,而 $\hat{\gamma}_0$ 是基于完整选项而得出的系数估计值的向量,且 $\hat{V}_1 - \hat{V}_0$ 是它们协方差矩阵的各自的估计值,那么统计值 $(\hat{\gamma}_1 - \hat{\gamma}_0)'(\hat{V}_1 - \hat{V}_0)(\hat{\gamma}_1 - \hat{\gamma}_0)$ 在 IIA 的限制是有效的这个零假设下,服从 $\chi^2(S)$分布。

如果这个零假设不被接受,那么一种可选择的方法是把方程 3.1 中的随机效用函数确定为多变量 probit 模型,$U_j = \sum_{s=1}^{S} \gamma_s W_{js} + \varepsilon_j \quad \varepsilon_j \sim N(0, \Sigma)$。其中,因为具有一个非纯量协方差矩阵 Σ,误差项就不再需要是独立的了。这个模型的问题是对多重正态积分和非限制协方差矩阵的估计进行计算这个实践问题(Greene, 2000:865)。

另一个放宽 IIA 限制的方法是对可供选择的选项进行分组,以使方差在组间不同,但在组内相同。这样就可以在选项组间放宽 IIA 假设,但是在每一组内仍然保留。这就是嵌套 logit 方法。这个问题不在这里讨论,详情及其应用请参考格林(1995)的论述。

第 4 章 | STATA 程序列表

简介

本章列出了产生前文所讨论的结果的 STATA 程序清单。在概论部分作者已经讲过——但值得重复——至少有四种众所周知的具高度好评的软件，除了其他方面，还解决与本书所讨论的类似的问题：SAS，SPSS v10.0（印第安纳大学统计与数学计算中心为在次序与非次序结果事件的分析中应用 SAS 和 SPSS 提供了很好的入门课程[65]），LIMDEP（Greene，1995），以及 STATA（STATA，1999）。命运的是，碰巧作者对 STATA 最熟悉。因此，这一章中的程序均用 STATA 编写。

下一部分包含了在第 2 章次序 logit 和 probit 模型中用到的程序清单，随后也包含了在第 3 章多类别 logit 模型中用到的程序清单。这些程序逐字地与相应的注释一起列出。这些注释的目的有三个：

(1) 解释特定的程序编码用来做什么（或做了什么）；

(2) 把该编码与前面内容涉及的某个特定的表格联系起来，以使读者看到那些结果是如何产生的；

(3) 把该编码与前面内容涉及的某些特定的方程联系起来，以使读者看到用以产生某个特定结果的方法。

次序 probit 和 logit 程序

```
version 6.0 /* 使用 STATA 版本 6.0 */;
```

use C:\SAGE\NLdta / * 读取 STATA 格式的数据:58 个变量，13164 个观察值 * /;

#delimit; / * 命令将由;终止 * /;

/ * 题目:使用剥夺分析例子的次序 logit 和 probit * /;

gen pnum = _N; / * 对每个人分配一个数字 * /;

/ * 对次序 logit，y 是因变量;y = 1 为没有被剥夺;y = 2 为轻度被剥夺;y = 3 为严重被剥夺 * /;

tabulate y rc，col; tabulate y sex，col;

/ * 用整个次样本估计 Oprobit 方程:表 2.2 * /;

oprobit y sex ct age age2 ret inac ue highed mided hnum snpar ard dwn crk ant col arm ban dry frm，table;

predict p1 p2 p3; / * 估计到的概率储存为 p1，p2，p3 * /;

sort pnum; / * 观察值用个人的数字以升序的形式分类 * /;

list pnum p1 p2 p3 in 1/25，noobs; / * 用个人数字列出前 25 个人的估计到的概率:表 2.4 * /;

summarize p1 p2 p3，detail; / * 详细概括估计到的个人概率:表 2.6 * /;

predict z，xb; / * 对每个人计算 Z 值 * /;

summarize z，mean; gen zm = r(mean); / * 计算 Z 的均值并存于 zm 这个变量中 * /;

/ * b[_cut1] and b[_cut2]储存了临界值 δ_1 和 δ_2 的估计值 * /;

gen p1 = normprob(_b[_cut1] − z); gen p2 = normprob(_b[_cut2] − z) − normprob(_b[_cut1] − z); gen p3 = 1 − normprob(_b[_cut2] − z);

/ * 对每个人 p1，p2 和 p3 为用方程 2.5a 至 2.5c 计算出来的概率的估计值,这些值与之前计算出来的 p1 − p3 的估计值相同,上面 * /;

gen q1 = normprob(_b[_cut1] − zm); gen q2 = normprob(_b[_cut2] − zm) − normprob(_b[_cut1] − zm); gen q3 = 1 − normprob(_b[_cut2] − zm);

/ * q1，q2 和 q3 是在设定取值等于样本均值的情况下计算出来的概率估计值:方程 2.18a 至 2.18c * /;

summarize p1 p2 p3; summarize q1 q2 q3; /* 比较概率:表 2.8 */;

drop p1 p2 p3 q1 q2 q3; /* 因为随后还要用到,在此放弃这些变量名 */

/* 计算边际效应:连续变量(Age) */;

gen a1 = normd(_b[_cut1] − z) * _b[age];

gen a2 = (normd(_b[_cut2] − z) − normd(_b[_cut1] − z) * _b[age];

gen a3 =− 1 * normd(_b[_cut2] − z) * _b[age];

/* a1 − a3 是对每个人计算的 AGEi 的边际效应:方程 2.13a 至 2.13c */;

gen am1 = normd(_b[_cut1] − zm) * _b[age]; gen am2 = (normd(_b[_cut2] − zm) − normd(_b[_cut1] − zm)) * _b[age];

gen am3 =− 1 * normd(_b[_cut2] − zm) * _b[age];

/* am1 − am3 是设定变量值等于样本均值时 AGE 的边际效应 */;

gen b1 = normd(_b[_cut1] − z) * _b[age2];

gen b2 = (normd(_b[_cut2] − z) − normd(_b[_cut1] − z)) * _b[age2];

gen b3 =− 1 * normd(_b[_cut2] − z) * _b[age2];

/* b1 − b3①是对每个人计算的 AGE_i^2 的边际效应:方程 2.13a 至 2.13c */;

gen bm1 = normd(_b[_cut1] − zm) * _b[age2];

gen bm2 = (normd(_b[_cut2] − zm) − normd(_b[_cut1] − zm)) * _b[age2];

gen bm3 =− 1 * normd(_b[_cut2] − zm) * _b[age2];

/* bm1 − bm3②是设定变量值等于样本均值时 AGE^2 的边际效应 */;

gen c1 = a1 + b1; gen c2 = a2 + b2; gen c3 = a3 + b3; gen cm1 = am1 + bm1; gen cm2 = am2 + bm2; gen cm3 = am3 + bm3;

/* 加上 AGE 和 AGE^2 的作用 */;

summarize c1 c2 c3; summarize cm1 cm2 cm3; /* 表 2.10 */;

---

① 原文中为 a1 − a3,其实有误。——译者注

② 原文中为 am1 − am3,其实有误。——译者注

drop z zm c1 a1 b1 c2 a2 b2 c3 a3 b3 cm1 am1 bm1 cm2 am2 bm2 cm3 am3 bm3;

/ * 计算边际效应:虚拟变量(Religion) * /;

gen cto = ct; / * 保存原值 * /;

replace ct = 1; / * 每个人都是天主教徒 * /;

predict p1 p2 p3; / * 当 CTi = 1 对每个人估计概率 * /;

predict z, xb; summarize z; gen zm = r(mean); / * 计算 Z 的均值并存于 zm 这个变量中 * /;

gen q1 = normprob(_b[_cut1] - zm); gen q2 = normprob(_b[_cut2] - zm) - normprob(_b[_cut1] - zm); gen q3 = 1 - normprob(_b[_cut2] - zm);

/ * q1, q2 和 q3 是在设定非宗教取值等于样本均值的情况下计算出来的概率估计值 * /;

summarize p1 p2 p3; summarize q1 q2 q3; / * 表 2.12 * /;

drop p1 p2 p3 q1 q2 q3 z zm;

replace ct = 0; / * 每个人都是新教徒 * /;

predict p1 p2 p3; / * 当 CTi = 0 对每个人估计概率 * /;

predict z, xb; summarize z; gen zm = r(mean); / * 计算 Z 的均值并存于 zm 这个变量中 * /;

gen q1 = normprob(_b[_cut1] - zm); gen q2 = normprob(_b[_cut2] - zm) - normprob(_b[_cut1] - zm); gen q3 = 1 - normprob(_b[_cut2] - zm);

/ * q1, q2 和 q3 是在设定非宗教取值等于样本均值的情况下计算出来的概率估计值 * /;

summarize p1 p2 p3; summarize q1 q2 q3; / * 表 2.12 * /;

drop p1 p2 p3 q1 q2 q3 z zm; replace ct = cto; / * 恢复原值 * /;

/ * 用整个次样本估计 Oprobit 方程:表 2.1 * /;

ologit y sex ct age age2 ret inac ue highed mided hnum snpar ard dwn crk ant col arm ban dry frm, table;

predict p1 p2 p3; /* 估计到的概率储存为 p1, p2, p3 */;

sort pnum; /* 观察值用个人的数字以升序的形式分类 */;

list pnum p1 p2 p3 in 1/25, noobs; /* 用个人数字列出前 25 个人的估计到的概率:表 2.4 */;

summarize p1 p2 p3, detail; /* 详细概括估计到的个人概率:表 2.6 */;

predict z, xb; /* 对每个人计算 Z 值 */;

summarize z, mean; gen zm = r(mean); /* 计算 Z 的均值并存于 zm 这个变量中 */;

/* b[_cut1] 和 b[_cut2] 储存了临界值 δ_1 和 δ_2 的估计值 */;

gen p1 = 1/(1 + exp(z - _b[_cut1])); gen p2 = 1/(1 + exp(z - _b[_cut2])) - 1/(1 + exp(z - _b[_cut1])); gen p3 = 1 - 1/(1 + exp(z - _b[_cut2]));

/* 对每个人 p1, p2 和 p3 为用方程 2.5a 至 2.5c 计算出来的概率的估计值,这些值与之前计算出来的 p1 - p3 的估计值相同,上面 */;

gen q1 = 1/(1 + exp(zm - _b[_cut1])); gen q2 = 1/(1 + exp(zm - _b[_cut2])) - 1/(1 + exp(zm - _b [_cut1])); gen q3 = 1 - 1/(1 + exp(zm - _b[_cut2]));

/* q1, q2 和 q3 是在设定取值等于样本均值的情况下计算出来的概率估计值:方程 2.18a 至 2.18c */;.

summarize p1 p2 p3; summarize q1 q2 q3; /* 比较概率:表 2.7 */;

drop p1 p2 p3 q1 q2 q3; /* 因为随后还要用到,在此放弃这些变量名 */;

/* 计算边际效应:连续变量(Age) */;

gen lden1 = 1/(1 + exp(z - _b[_cut1]));

gen lden2 = 1/(1 + exp(z - _b[_cut2]));

gen lden3 = 1/(1 + exp(z - _b[_cut2]));

/* logit 概率计算出来了:方程 2.9a 至 2.9c */;

gen a1 = lden1 * (1 - lden1) * _b[age]; gen b1 = lden1 * (1 - lden1) * _b[age2]; gen c1 = a1 + b1; gen a2 = (lden2 * (1 - lden2) - lden1 *

```
(1 — lden1)) * _b[age];
gen b2 = (lden2 * (1 — lden2) — lden1  * (1 — lden1)) * _b[age2];
gen c2 = a2+b2;
gen a3 =— lden3 * (1 — lden3) *  _b[age];
gen b3 =— lden3 * (1 — lden3) * _b[age2];
gen c3 = a3 + b3;
/*  a1 — a3 是对每个人计算的 AGEi 的边际效应:方程 2.12a 至 2.12c */;
/*  b1 — b3 是对每个人计算的 AGEi^2 的边际效应:方程 2.12a 至2.12c */;
/* c1 — c3 是这些效应的和 */;
replace lden1 = 1/(1 + exp(zm — _b[_cut1]));
replace lden2 = 1/(1 + exp(zm — _b[_cut2]));
replace lden3 = 1/(1 + exp(zm — _b[_cut2]));
/*  logit 概率用均值计算 */;
gen am1 = lden1 * (1 — lden1) * _b[age];
gen bm1 = lden1 * (1 — lden1) * _b[age2];
gen cm1 = am1 + bm1;
gen am2 = (lden2 * (1 — lden2) — lden1 * (1 — lden1)) * _b[age];
gen bm2 = (lden2 * (1 — lden2) — lden1 * (1 — lden1)) * _b[age2];
gen cm2 = am2 + bm2;
gen am3 =— lden3 * (1 — lden3) * _b[age];
gen bm3 =— lden3 * (1 — lden3) * _b[age2];
gen cm3 = am3 + bm3;
```

/* am1 — am3[①]是用样本均值计算的 AGEi 的边际效应 */; /* bm1 — bm3[②]是用样本均值计算的 AGE 的边际效应 */; /* cm1 — cm3[③]是这些效应

① 原为 a1 — a3,其实有误。——译者注

② 原为 b1 — b3,其实有误。——译者注

③ 原为 c1 — c3,其实有误。——译者注

的和 * /;

summarize c1 c2 c3; summarize cm1 cm2 cm3; / * 表 2.9 * /;

drop z zm c1 a1 b1 c2 a2 b2 c3 a3 b3 cm1 am1 bm1 cm2 am2 bm2 cm3 am3 bm3;

replace ct = 1; / * 每个人是天主教徒 * /;

predict p1 p2 p3; / * 当 CTi = 1 对每个人估计概率 * /; predict z, xb; summarize z; gen zm = r(mean); / * 计算 Z 的均值并存于 zm 这个变量中 * /;

gen q1 = 1/(1 + exp(zm — _b[_cut1])); gen q2 = 1/(1 + exp(zm — _b[_cut2])) — 1/(1 + exp(zm — _b[_cut1])); gen q3 = 1 — 1/(1+exp(zm — _b[_cut2]));

/ * q1, q2 和 q3 是在设定非宗教取值等于样本均值的情况下计算出来的概率估计值 * /;

summarize p1 p2 p3; summarize q1 q2 q3; / * 表 2.11 * /;

drop p1 p2 p3 q1 q2 q3 z zm;

replace ct = 0; / * 每个人都是新教徒 * /;

predict p1 p2 p3; / * 当 CTi = 0 对每个人估计概率 * /;

predict z xb; summarize z; gen zm = r(mean); / * 计算 Z 的均值并存于 zm 这个变量中 * /;

gen q1 = 1/(1 + exp(zm — _b[_cut1])); gen q2 = 1/(1 + exp(zm — _b[_cut2])) — 1/(1 + exp(zm — _b[_cut1])); gen q3 = 1 — 1/(1 +exp(zm — _b[_cut2]));

/ * q1, q2 和 q3 是在设定非宗教取值等于样本均值的情况下计算出来的概率估计值 * /;

summarize p1 p2 p3; summarize q1 q2 q3; / * 表 2.11 * /;

drop p1 p2 p3 q1 q2 q3 z zm;

replace ct = cto; / * 恢复原值 * /;

/ * 用新教徒次样本估计 Oprobit 方程 * /;

ologit y sex age age2 ret inac ue highed mided hnum snpar ard dwn crk ant

col arm ban dry frm if ct = 0, table;

lrtest,saving(0); /*为 LR 检验储存概率值*/;

ologit y sex age age2 ret inac ue highed mided hnum snpar ard crk ant col arm ban frm if ct = 0, table;

lrtest; /* LR 检验 dwn 和 dry 的系数联合等于 0 */;

/*对新教徒用新教徒的系数进行预测,且计算均值*/;

predict ppa1 ppa2 ppa3 if ct = 0; egen mppa1 = mean(ppa1); egen mppa2 = mean(ppa2); egen mppa3 = mean(ppa3);

predict pps1 pps2 pps3 if ct = 0 & snpar = 1;

egen mpps1 = mean(pps1); egen mpps2 = mean(pps2); egen mpps3 = mean(pps3); predict ppr1 ppr2 ppr3 if ct = 0 & ret = 1; egen mppr1 = mean (ppr1); egen mppr2 = mean(ppr2); egen mppr3 = mean(ppr3);

predict ppi1 ppi2 ppi3 if ct = 0 & inac = 1;

egen mppi1 = mean(ppi1); egen mppi2 = mean(ppi2); egen mppi3 = mean(ppi3); predict ppu1 ppu2 ppu3 if ct = 0 & ue = 1; egen mppu1 = mean (ppu1); egen mppu2 = mean(ppu2); egen

mppu3 = mean(ppu3);

predict ppl1 ppl2 ppl3 if ct = 0 & resnum = 1;

egen mppl1 = mean(ppl1); egen mppl2 = mean(ppl2); egen mppl3 = mean(ppl3);

/*用天主教徒次样本估计 Oprobit 方程*/;

ologit y sex age age2 ret inac ue highed mided hnum snpar ard dwn crk ant col arm ban dry frm if ct = 1, table;

lrtest, saving(0); /*为 LR 检验储存概率值*/;

ologit y sex age age2 ret inac ue highed mided hnum snpar ard crk ant col arm ban frm if ct = 1, table; lrtest; /* LR 检验 dwn 和 dry 的系数联合等于 0 */;

/*对天主教徒用天主教徒的系数进行预测,且计算均值*/;

predict cca1 cca2 cca3 if ct = 1;

egen mcca1 = mean(cca1); egen mcca2 = mean(cca2); egen mcca3 = mean(cca3); predict ccs1 ccs2 ccs3 if ct = 1 & snpar = 1; egen mccs1 = mean (ccs1); egen mccs2 = mean(ccs2); egen mccs3 = mean(ccs3);

predict ccr1 ccr2 ccr3 if ct = 1 & ret = 1;

egen mccr1 = mean(ccr1); egen mccr2 = mean(ccr2); egen mccr3 = mean(ccr3); predict cci1 cci2 cci3 if ct = 1 & inac = 1; egen mcci1 = mean (cci1); egen mcci2 = mean(cci2); egen mcci3 = mean(cci3);

predict ccu1 ccu2 ccu3 if ct = 1 & ue = 1;

egen mccu1 = mean(ccu1); egen mccu2 = mean(ccu2); egen mccu3 = mean(ccu3);

predict cel1 cel2 cel3 if ct = 1 & resnum = 1;

egen mcel1 = mean(ccl1); egen mcel2 = mean(cel2); egen mcel3 = mean (cel3);

/*对新教徒用天主教徒的系数进行预测,且计算均值*/;

predict pca1 pca2 pca3 if ct = 0;

egen mpca1 = mean(pca1); egen mpca2 = mean(pca2); egen mpca3 = mean(pca3);

predict pcs1 pcs2 pcs3 if ct = 0 & snpar = 1;

egen mpcs1 = mean(pcs1); egen mpcs2 = mean(pcs2); egen mpcs3 = mean(pcs3);

predict pcr1 pcr2 pcr3 if ct = 0 & ret = 1;

egen mpcr1 = mean(pcr1); egen mpcr2 = mean(pcr2); egen mpcr3 = mean(pcr3);

predict pci1 pci2 pci3 if ct = 0 & inac = 1;

egen mpci1 = mean(pci1); egen mpci2 = mean(pci2); egen mpci3 = mean(pci3);

predict pcu1 pcu2 pcu3 if ct = 0 & ue = 1;

egen mpcu1 = mean(pcu1); egen mpcu2 = mean(pcu2); egen mpcu3 =

```
mean(pcu3);
    predict pcl1 pcl2 pcl3 if ct = 0 & resnum = 1;
    egen mpcl1 = mean(pcl1); egen mpcl2 = mean(pcl2); egen mpcl3 =
mean(pcl3);
    /*现在百分比贡献(percentage contributions)可以计算出来了*/;
    gen Aa1 = [(mppa1 - mpca1)/(mppa1 - mcca1)] * 100;
    gen Aa2 = [(mppa2 - mpca2)/(mppa2 - mcca2)] * 100;
    gen Aa3 = [(mppa3 - mpca3)/(mppa3 - mcca3)] * 100;
    gen As1 = [(mpps1 - mpcs1)/(mpps1 - mccs1)] * 100;
    gen As2 = [(mpps2 - mpcs2)/(mpps2 - mccs2)] * 100;
    gen As3 = [(mpps3 - mpcs3)/(mpps3 - mccs3)] * 100;
    gen Ar1 = [(mppr1 - mpcr1)/(mppr1 - mccr1)] * 100
    gen Ar2 = [(mppr2 - mpcr2)/(mppr2 - mccr2)] * 100;
    gen Ar3 = [(mppr3 - mpcr3)/(mppr3 - mccr3)] * 100;
    gen Ai1 = [(mppi1 - mpci1)/(mppi1 - mcei1)] * 100;
    gen Ai2 = [(mppi2 - mpci2)/(mppi2 - mcci2)] * 100;
    gen Ai3 = [(mppi3 - mpci3)/(mppi3 - mcci3)] * 100;
    gen Au1 = [(mppu1 - mpcu1)/(mppu1 - mccu1)] * 100;
    gen Au2 = [(mppu2 - mpcu2)/(mppu2 - mccu2)] * 100;
    gen Au3 = [(mppu3 - mpcu3)/(mppu3 - mccu3)] * 100;
    gen Al1 = [(mppl1 - mpel1)/(mppl1 - mcel1)] * 100;
    gen Al2 = [(mppl2 - mpel2)/(mppl2 - mcel2)] * 100;
    gen Al3 = [(mppl3 - mpel3)/(mppl3 - mcel3)] * 100;
    summarize
    ppa1 ppa2 ppa3 pca1 pca2 pca3 cca1 cca2 cca3 pps1 pps2 pps3 pcs1 pcs2
pcs3 ccs1 ccs2 ccs3 ppr1 ppr2 ppr3 pcr1 pcr2 pcr3 ccr1 ccr2 ccr3 ppi1 ppi2 ppi3
pci1 pci2 pci3 cci1 cci2 cci3 ppu1 ppu2 ppu3 pcu1 pcu2 pcu3 ccu1 ccu2 ccu3 ppl1
ppl2 ppl3 pcl1 pcl2 pcl3 ccl1 ccl2 ccl3;
```

/ * 表 2.16 * /;

summarize Aa1 Aa2 Aa3 As1 As2 As3 Ar1 Ar2 Ar3 Ai1 Ai2 Ai3 Au1 Au2 Au3 Al1 Al2 Al3; / * 表 2.17 * /;

/ * 程序结束 * /;

多类别 logit 程序

use c:\SAGE\GB.dta / * 读取 STATA 格式的数据:24 个变量,98732 个观察值:这些数据为白人、黑人和印第安人全职男性雇员,25—45 岁 * /;

delimit;

/ * 题目:使用职业类别分析例子的多类别 logit * /;

gen pnum = _N; / * 对每个人分配一个数字 * /;

/ * 多类别 logit 中 y 是因变量,且

y = 1 如果这个人受雇于非技术/半技术工人职业类别

y = 2 如果这个人受雇于有技术体力/非体力劳动者职业类别

y = 3 如果这个人受雇于专业/管理/技术人员职业类别 * /;

/ * 对所有人、白人、黑人、印第安人列表 * /;

tab y; tab y if blk = 0 &ind = 0; tab y if blk = 1; tab y if ind = 1;

/ * 现在估计多类别 logit 方程,表 3.2:模型设定见方程 3.14。注:以下 base(1)设结果 1 为基准值 * /;

mlogit y age age2 blk blkovs ind indovs ovsbn north indnth

blknth south indsth blksth highed indhe blkhe mided

indme blkme subbus subsei bush scih, base(1);

lrtest, saving(0); / * 为 LR 检验储存似然对数 * /;

/ * 确定零限制:对 y = 2,blk 和 blkovs 从方程中去掉;对 y = 3,ind, blknth, scih, bush, subsci 从方程中去掉;indme, indhe, blkme 在两个方程中都去掉 * /;

constraint define 1 [2]blkovs = 0; constraint define 2 [3]blk = 0;

constraint define 3 [3]ind = 0; constraint define 4 [3]blknth = 0;

constraint define 5 [3]scih = 0; constraint define 6 [3]bush = 0;

constraint define 7 [3]subsci = 0; constraint define 8 indhe;

constraint define 9 indme; constraint define 10 blkme;

/*现在在强加限制 1—10 的情况下估计 mlogit 方程,表 3.3。注:下面的 constr(1—10)为强加限制的命令*/;

mlogit y age age2 blk blkovsind indovs ovsbn north indnth

blknth south indsth blksth highed indhe blkhe mided indme

blkme subbus subsci bush scih, constr(1—10) base(1);

lrtest; /*似然比检验完成:chi2(13) = 8.1 零限制没有被拒绝。*/;

/*对每个人对结果 1, 2, 3 估计到的概率分别储存为 p1, p2, p3 */;

predict p1, outcome(1); predict p2, outcome(2); predict p3, outcome(3);

/*对所有人,白人,黑人和印第安人,概括概率估计值(表 3.4 上部分)*/;

summarize p1 p2 p3; summarize p1 p2 p3 if blk. = 0 & ind = 0; summarize p1 p2 p3 if blk = 1; summarize p1 p2 p3 if ind = 1;

drop p1 p2 p3; /*因为随后还要用到,在此放弃这些变量名*/;

/*对每个人对结果 1, 2, 3 计算 Z 值,并储存为 z1, z2, z3 */;

predict z1, outcome(1) xb; predict z2, outcome(2) xb; predict z3, outcome(3) xb;

/*z1, z2, z3 的均值储存为 zm1, zm2, zm3 */;

summarize z1, mean; gen zm1 = r(mean); summarize z2, mean;

gen zm2 = r(mean); summarize z3, mean; gen zm3 = r(mean);

/*对所有人,用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率*/;

gen sum = 1 + exp(zm2) + exp(zm3); gen p1 = 1/sum;

gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum; summarize p1 p2 p3;

/*对所有人展示了概率估计值(表 3.4 下部分)*/;

drop sum p1 p2 p3 zm1 zm2 zm3; /*因为随后还要用到,在此放弃这些变量名*/;

/ * 现在对白人计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 * /;

summarize z1 if blk = 0 & ind = 0, mean; gen zm1 = r(mean); summarize z2 if blk = 0 & ind = 0, mean; gen zm2 = r(mean); summarize z3 if blk = 0 & ind = 0, mean; gen zm3 = r(mean);

/ * 对白人,用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率 * /;

gen sum = 1+exp(zm2)+exp(zm3); gen p1 = 1/sum; gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum; summarize p1 p2 p3 if blk = 0 &ind = 0;

/ * 对白人展示了概率估计值(表 3.4 下部分) * /;

drop sum p1 p2 p3 zm1 zm2 zm3; / * 因为随后还要用到,在此放弃这些变量名 * /;

/ * 现在对黑人计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 * /;

summarize z1 if blk = 1, mean; gen zm1 = r(mean);

summarize z2 if blk = 1, mean; gen zm2 = r(mean);

summarize z3 if blk = 1, mean; gen zm3 = r(mean);

/ * 对黑人,用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率 * /;

gen sum = 1+exp(zm2)+exp(zm3); gen p1 = 1/sum; gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum;

summarize p1 p2 p3 if blk = 1; / * 对黑人展示了概率估计值(表 3.4 下部分) * /;

drop sum p1 p2 p3 zml zm2 zm3; / * Variables are released for subsequent use * /;

/ * 现在对印第安人计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 * /;

summarize z1 if ind = 1, mean; gen zm1 = r(mean);

summarize z2 if ind = 1, mean; gen zm2 = r(mean);

summarize z3 ifind = 1, mean; gen zm3 = r(mean);

/ * 对印第安人,用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率 * /;

gen sum = 1 +exp(zm2)+exp(zm3); gen p1 = 1/sum; gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum;

summarize p1 p2 p3 if ind = 1; /＊对印第安人展示了概率估计值(表 3.4 下部分)＊/;

drop sum p1 p2 p3 zm1 zm2 zm3; drop z1 z2 z3; /＊因为随后还要用到,在此放弃这些变量名＊ /;

/＊接着为种族模拟分析＊ /;

gen blko = blk; gen indo = ind; /＊保存种族变量＊/;

replace blk = 0; replace ind = 0; /＊每个人都是白人＊ /;

/＊所有涉及 blk 和 ind 的交互变量都需要重新计算＊ /;

replace indnth = ind＊north; replace indsth = ind＊south;

replace blknth = blk＊north; replace blksth = blk＊south;

replace blkovs = blk＊ovsbn; replace indovs = ind＊ovsbn;

replace blkhe = blk＊highed;

/＊对每个人对结果 1, 2, 3 估计到的概率分别储存为 p1, p2, p3＊/;

predict p1, outcome(1); predict p2, outcome(2); predict p3, outcome (3); summarize p1 p2 p3;

/＊对整个样本概括概率估计值:每个人都假定为白人(表 3.5 上部分)＊/;

drop p1 p2 p3;

/＊对每个人对结果 1, 2, 3 计算 Z 值并储存于 z1, z2, z3: 每个人都假定为白人＊ /;

predict z1, outcome(1) xb; predict z2, outcome(2) xb; predict z3, outcome(3) xb;

/＊对整个样本,计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 ＊/;

summarize z1, mean; gen zm1 = r(mean); summarize z2, mean; gen zm2 = r(mean); summarize z3, mean; gen zm3 = r(mean);

/＊用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率:每个人都假定为白人＊ /;

gen sum = 1 + exp(zm2) + exp(zm3); gen p1 = 1Isum; gen p2 = exp (zm2)/sum; gen p3 = exp(zm3)/sum;

summarize p1 p2 p3; / * 对整个样本概括概率估计值:每个人都假定为白人(表 3.5 下部分) * /;

drop z1 z2 z3 sum p1 p2 p3 zm1 zm2 zm3; / * 因为随后还要用到,在此放弃这些变量名 * /;

replace blk = 1; / * 每个人都是黑人 * /;

/ * 所有涉及 blk 和 ind 的交互变量都需要重新计算 * /;

replace blknth = blk * north; replace blksth = blk * south;

replace blkovs = blk * ovsbn; replace blkhe = blk * highed;

/ * 没有必要替换 ind * 变量因为 ind 已经等于 0 * /;

/ * 对每个人对结果 1, 2, 3 估计到的概率分别储存为 p1, p2, p3 * /;

Predict p1, outcome(1); predict p2, outcome(2); predict p3, outcome(3);

summarize p1 p2 p3; drop p1 p2 p3; / * 对整个样本概括概率估计值:每个人都假定为黑人(表 3.5 上部分) * /;

/ * 对每个人对结果 1, 2, 3 计算 Z 值并储存于 z1, z2, z3: 每个人都假定为黑人 * /;

predict z1, outcome(1) xb; predict z2, outcome(2) xb; predict z3, outcome(3) xb;

/ * 对整个样本,计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 * /;

summarize z1, mean; gen zm1 = r(mean); summarize z2, mean; gen zm2 = r(mean); summarize z3, mean; gen zm3 = r(mean);

/ * 用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率:每个人都假定为黑人 * /;

gen sum = 1 + exp(zm2) + exp(zm3); gen p1 = 1/sum;

gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum;

summarize p1 p2 p3; / * 对整个样本概括概率估计值:每个人都假定为黑人(表 3.5 下部分) * /;

drop z1 z2 z3 sum p1 p2 p3 zm1 zm2 zm3; / * 因为随后还要用到,在此放弃这些变量名 * /;

replace ind = 1; replace blk = 0; /* 每个人都是印第安人 */;

/* 所有涉及 blk 和 ind 的交互变量都需要重新计算 */;

replace indnth=ind * north; replace indsth=ind * south;

replace blknth=blk * north; replace blksth=blk * south;

replace blkovs=blk * ovsbn; replace indovs=ind * ovsbn;

replace blkhe=blk * highed;

/* 对每个人对结果 1, 2, 3 估计到的概率分别储存为 p1, p2, p3 */;

predict p1, outcome(1); predict p2, outcome(2); predict p3, outcome(3);

summarize p1 p2 p3; drop p1 p2 p3; /* 对整个样本概括概率估计值:每个人都假定为印第安人(表 3.5 上部分) */;

/* 对每个人对结果 1, 2, 3 计算 Z 值并储存于 z1, z2, z3: 每个人都假定为印第安人 */;

predict z1, outcome(1) xb; predict z2, outcome(2) xb; predict z3, outcome(3) xb;

/* 对整个样本,计算 z1 z2 z3 的均值,并储存于 zm1, zm2, zm3 */;

summarize z1, mean; gen zm1 = r(mean); summarize z2, mean; gen zm2 = r(mean); summarize z3, mean; gen zm3 = r(mean);

/* 用方程 3.12a 和 3.12b 计算结果 1, 2, 3 的概率:每个人都假定为印第安人 */;

gen sum = 1 + exp(zm2) + exp(zm3); gen p1 = 1/sum;

gen p2 = exp(zm2)/sum; gen p3 = exp(zm3)/sum;

summarize p1 p2 p3; /* 对整个样本概括概率估计值:每个人都假定为印第安人表 3.5 下部分 */;

drop z1 z2 z3 sum p1 p2 p3 zm1 zm2 zm3; /* 因为随后还要用到,在此放弃这些变量名 */;

replace blk = blko; replace ind = indo; /* 恢复原来的变量 */;

/* 程序结束 */;

注释

[1] 在网站 http://www.indiana.edu/~statmath/stat/all/cat/giant.html 可以找到。

[2] 只有两种结果时，次序 logit 与 probit 这两种方法都可以使用，不管因变量是有序的还是非有序的。

[3] 但是，有序模型可以基于除 logit 或 probit 分布之外的其他分布。比如，对偏斜有序数据(skewed ordinal data)的双对数(log-log)、负双对数(negative log-log)或互补双对数(complementary log-log)分布(见 Agresti 1990)。大致要点是，次序 logit 和 probit 模型是一系列更广泛得多的处理有序因变量的观点中的一部分。其实，如果结果的数目很大(比如，大于 20)，那么次序 logit 和 probit 的方法就会变得很笨重，用其他方法可能会更合适(关于这点的讨论，见 Jöreskog and Sörbom, 1988:44—45，及 1993:1—17)。

[4] 排序可以基于收入，这种情况下职业具有一个客观的等级；或者也可基于文化假设和偏见。

[5] 比"把结果当成是有序的，除非我们有好的理由不强加一个排序"更好。

[6] 记住 $N = N_1 + N_2 + N_3$。

[7] 因为它们的计算方法，这些估计值被称为最大似然估计值(maximum likelihood estimates)。

[8] 还有别的把有序结果模型化的方法，比如相邻类别(adjacent categories)logit 模型，构造型(stereotype)模型，logit 模型和连续比(continuation ratio)logit 模型(见 Agresti, 1996:216—220)。

[9] 它非常类似于一个自由度为 7 的 t 分布(Greene, 2000:815)。

[10] 关于这个属性的讨论见 Brant(1990)。

[11] 其实 Greene (2000)把它的临界值记为从 0 开始，所以在它的注释中，$\mu = 0$。但为了方便比较，我把他的临界值起点记为 1。

[12] 记住 Greene(2000:876)把 $\hat{W}_i$ 写成$\beta' x$。

[13] 因为 $\mu_1 = 0$, $\delta_1 = -\beta_0$，所以 STATA 的第一个临界值是负的截距。

[14] 均值为 0 方差为 1 的正态分布。

[15] 记住这三个概率之和必须为 1。

[16] 与其中一个结果相关的系数必须为了可识别性而被标准化。

[17] 这些人相当于这次人口普查记录的一个 2%样本。对这个数据的更完整的描述见 Borooah (2000a)。

[18] 某个人被认为没有被剥夺如果 $D_i = 0$，轻度被剥夺如果 $0 < D_i \leqslant \bar{D}$，严重被剥夺如果 $D_i > \bar{D}$，其中 $\bar{D}$ 为剥夺指数的均值。

[19] 另一种标准化，设定 17 的人 $AGE_i = 1$，18 岁的 $AGE_i = 2$，等等。

[20] 这些是通常在 18 岁或 18 岁以上获得的文凭。

[21] 他们是：Belfast(地区 1)；Ards，Castlereagh，North Down(地区 2)；Down，Lisburn(地区 3)；Carrickfergus，Larne，Newtwoenabbey(地区 4)；Antrim，Ballymena and Ballymoney(地区 5)；Armagh，Newry & Mourne(地区 7)；Coleraine，Cookstown，Maghrafelt，Moyle(地区 6)；Banbridge，Craigavon，Dungannon(地区 8)；Derry，Limavady(地区 9)；and Fermanagh，Omagh，Strabane(地区 10)。因为多重共线性，不能 10 地区全都放在方程 2.16 中。$AREA_1$(Belfast)这个地区从方程中去掉了。

[22] 所有非罗马天主教徒识别为"基督徒"，虽然后者还包括要么没有陈述宗教信仰(北爱尔兰居民的 7.2%)，要么宣称他们没有信仰的人(北爱尔兰居民的 3.8%)。

[23] 注意(a)在 STATA 里，$\beta_1=0$，(b)从方程 2.16，$dD_i/dAGE_i=\beta_4+2\beta_5 AGE_i$。

[24] 这些比率表示沃尔德统计量(Wald statistic)的一种形式。

[25] 为了测量平行斜率假设，这个模型还用多类别 logit 进行了估计，产生了一个等于 −12380.36 的似然值(记为 L_2)。由 $2(L_2-L_1)$ 计算出来的值为 86.4(与次序 logit 模型进行的比较)，依照严格的似然比检验的解释，这个值超过了 $\chi^2(20)$ 的 5%临界值 31.41。但是，正如前面指出的，$2(L_2-L_1)$ 的值只是提示性的，因为这个统计值并没有提供似然比检验的基础。

[26] 说明：(表中的) $xb+u=$ (文字中的) $\hat{Z}_i+\varepsilon_i$。

[27] 临界点其实只是模型的系数，且与斜率系数一起用最大似然方法估计而得。

[28] 这些比例在这个章节开头的时候就设定了。

[29] 虽然当平均数被定义为个人概率的中位数(而不是均值)时结果会不一样。

[30] 注意在 STATA 中，$\beta_1=0$。

[31] 为了方便表述，随后的讨论全都针对 logit 模型。

[32] 比如，与 8%的新教徒相比，样本中将近 16%的天主教徒为单身父/母亲。

[33] 注意这些预测到的均值会与天主教徒与新教徒处于不同剥夺程度的样本比例相同。

[34] 即新教徒的属性用新教徒的系数进行评估。

[35] 在失业者和退休者的情况下，这些概率几乎不变。

[36] 在具有相同的处于轻度被剥夺的平均概率的意义上来说。

[37] 这些人相当于这次人口普查记录的一个 2%样本。对这个数据的更完整的描述见 Borooah (2000b)。

[38] 缺省值为没有 18 岁或 18 岁以上的文凭。

[39] 只在这个人具有 18 岁或 18 岁以上的文凭时有意义。缺省值为文科类课程。

[40] 缺省值为白人。

[41] 缺省值为伦敦。

[42] 关于英国的标准区域划分：中英格兰东(East Midlands)；东英格利亚(East Anglia)；东南部(the Southeast)(伦敦除外)；西南部(the Southwest)。

[43] 北部；约克郡(Yorkshire)；中英格兰西(West Midlands)；西北部；威尔士(Wales)；苏格兰(Scotland)。

[44] 注意这三种结果中的任何一种都可以被选择为基准结果。

[45] 当然，如果对某个特定组别相关系数为正的，那么它就会享受"种族的意外收获"。

[46] 这些比率表示 沃尔德统计量(Wald statistic)的一种形式。

[47] 除了年龄：年龄更大的人更加可能处于更高的职业类别，或者，就方程 3.14 的估计系数来说，$\hat{\theta}_{j1}>0,\ j=1,\ 2$。

[48] 除非有特殊说明相反情况，否则本文的讨论均带有其他条件相同这一条件。

[49] 虽然在 PMT 这个类别上对居住在北部的黑人并不适用。

[50] 或者说，样本中 41.5%的所有男性，41.8%的白人男性，29.3%的黑人男性，和 35.1%的印度男性处于 PMT 职业类别中。

[51] 通过把方程里面的 $\bar{Z}_j$，视情况而定，换成 $\bar{Z}_{Wj}$，$\bar{Z}_{bj}$ 和 $\bar{Z}_{Ij}$。

[52] 或者同等地利用样本白人中个人的决定变量 X_{ik} 的均值 $\bar{X}_{Wr}$ 进行计算。

[53] 应该注意的是，这一部分介绍的以模拟为基础的策略是与关于组间结果差别相关的一系列具体问题有关的。至于模拟策略的更加概括性的方法，请见 Tomz, Wittenberg and King(1999)。Tom 等人(1999)开发了一个程序 Clarify，这个程序利用随机模拟把原始的统计程序结果转换成

让研究者直接感兴趣的结果。这个程序是为了在 STATA 中运行而设计的。

[54] 类似的应用也可以在假定样本中每个人都为印度人的情况下实行。

[55] 其中 α_{jr} 是白人的系数，$\alpha_{jr}+\gamma_{jr}$ 是印度人的系数。

[56] $\hat{Z}_{ij}$ 是用表 3.3 中所示的估计值联合每个人的决定变量的值计算出来的。在方程 3.14 中，当所有人是白人时，$IND_i = BLK_i = 0$；当所有人是黑人时，$BLK_i = 1$。

[57] 另一方面，如果 $\bar{p}_j^B > \bar{p}_j^W$，则 $\lambda_j^B > 1$，存在种族优势。

[58] 另一方面，如果 $s_j^B > s_j^W$，则 $\mu_j^B > 1$，存在总体获利。

[59] 这仅仅取决于样本比例的比，而它当然是保持不变的。

[60] 在多类别模型中，风险比对数仅仅取决于与 j 和 k 相关的系数(间的差别)，与任何其他结果的系数无关。

[61] Domencich and McFadden(1996)。

[62] 这个不足之处是经济学中许多在经验上方便的函数形式共有的，比如柯布—道格拉斯(Cobb-Douglas)或者不变替代弹性(CES)(Constant Elasticity of Substitution)函数形式。

[63] 虽然必须强调，新选项的引进可能是对已有选项的分割：比如把熟练的体力劳动者/非体力劳动者分成熟练的体力劳动者和熟练的非体力劳动者两个选项。

[64] 见先前讨论的如何在引进一个新选项时重新计算概率。

[65] 在网站 http://www.indiana.edu/~statmath/stat/all/cat/giant.html 可以找到。

参考文献

AGRESTI, A (*1996*). *An introduction to categorical data analysis*. New York: John Wiley.

ALDRICH, J. H., and NELSON, F. D. (1984). Linear probability, logit, and probit models, *Quantitative Applications in the Social Sciences*, *07—045*. Beverly Hills, CA: Sage.

BEN-AKIVA, M., and LERMAN, S. (198S). *Discrete choice analysis*. London: MIT Press.

BOROOAH, V. K., and CARCACH, C. (1997). Fear and crime: Evidence from Australia. *British Journal of Criminology*, *37*, 634—656.

BOROOAH, V K. (2000). Targeting social need: Why are deprivation levels in Northern Ireland higher for Catholics than for Protestants? *Journal of Social Policy*, 29, 281—301.

BOROOAH, V K. (2001). How do employees of ethnic origin fare on the occupational ladder in Britain? *Scol/ish Journal of Political Economy*, *48*, 1—26.

BRANT, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrika*, *44*, 131—140.

CRAMER, J. (1999). Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society*, *Series D(The Statistician)*, *88*, 85—94.

DEMARIS, A(1992). Logit modelling. *Quantitative Applications in the Social Sciences*, *07—086*. Newbury Park, CA: Sage.

DESAI, M., and SHAH, A(1988). An econometric approach to the measurement of poverty. *Oxford Economic Papers*, *40*, 505—522.

DOMENCICH, T. A, and McFADDEN, D. (1996). *Urban travel demand: A behavioral analysis*. Amsterdam: North-Holland.

GREENE, W. H. (1995). *LIMDEP, version 7.0: User's manual*. Bellport, NY: Econometric Software.

GREENE, W. H. (2000). *Econometric analysis*. Englewood Cliffs, NJ: Prentice Hall(4th ed.,).

HAUSMAN, J., and McFADDEN, D. (1984). A specification test for the multinomial logit model. *Econometrica*, *52*, 1219—1240.

HENSHER, D. (1986). *Simultaneous estimation of hierarchical logit mode choice models* (Working Paper No. 24). MacQuarie University, School of Economic and Financial Studies.

JORESKOG, K., and SORBOM, D. (1988). *PRELIS: A program for multivariate data screening and data summarization*. Chicago: Scientific Software Inc.

JORESKOG, K., and SORBOM, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hilllsdale, NJ: Lawrence Erlbaum.

KAY, R., and LITTLE, S. (1986). Assessing the fit of the logistic model: A case study of children with haemolytic uraemic syndrome. *Applied Statistics*, *35*, 16—30.

LIAO, T. F. (1994). Interpreting probability models. *Quantitative Applications in the Social Sciences*, *07—101*. Newbury Park, CA: Sage.

McFADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York: Academic Press.

MENARD, S. (1995). Applied regression analysis. *Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

NOLAN, B., and WHELAN, C. T. (1996). *Resources deprivation and poverty*. Oxford: Clarendon Press.

PIACHAUD, D. (1987). Problems in the definition of poverty. *Journal of Social Policy*, *16*, 147—164.

SCHMIDT, P., and STRAUSS, R. P. (1975). The Prediction of occupation using multiple logit models. *International Economic Review*. *13*, 471—486.

STATA(1999). *Stata reference manual release* 6. College Station, TX: Stata Press.

TOMZ, M., WITTENBERG, J., and KING, G. (1999). *CLARIFY: Software for interpreting and presenting statistical results*, *version* 1. 2. 1. Cambridge, MA: Harvard University Press. (http://gking. harvard. edu/).

TOWNSEND, P. (1979). *Poverty in the United Kingdom*. Harmondsworth: Penguin.

VEALL, M. R., and ZIMMERMANN, K. F. (1996). Pseudo-R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, *10*, *241—260*.

译名对照表

adjacent categories logit	相邻类别 logit(模型)
base outcome	"基准"结果
best linear unbiased estimator(BLUE)	最优线性无偏估计
binary choice model	二分选择模型
binary model	二分模型
categorical variable	类别变量
Cobb-Douglas	柯布—道格拉斯(函数形式)
complementary log-log distribution	互补双对数分布
Constant Elasticity of Substitution(CES)	不变替代弹性(函数形式)
continuation ratio logit model	连续比 logit 模型
critical value	临界值
Cross-Price Elasticity of Demand	需求的交叉弹性
cumulative probability distribution	累积概率分布
cutoff parameter	临界参数
degree of freedom	自由度
discrete dependent variable	离散因变量
dummy variable	虚拟变量
error term	误差项
estimate	估计值
estimation	估计
explanatory power	解释力
generalized logit model	广义 logit 模型
goodness of fit test	拟合优度检验
Independence of Irrelevant Alternatives(IIA)	无关选项独立性
Independently and Identically Distributed(IID)	独立同分布
individual observations	个体观测值
interaction effect	交互作用
interaction variables	交互变量
intercept term	截距项
latent regression	潜在回归
latent variable	潜在变量
likelihood ratio	似然比
likelihood-ratio test	似然比检验
linear function	线性函数
log likelihood function	对数似然函数
logistically distributed	逻辑分布

log-log distribution	双对数分布
marginal effect	边际效应
maximum likelihood estimates	最大似然估计值
model specificaitn	模型设定
multinomial distribution	多项分布
multinomial logit/probit	多类别 logit/probit
multinormal integration	多重正态积分
multiple outcome model	多元结果模型
negative log-log distribution	负双对数分布
nonscalar covariance matrix	非纯量协方差矩阵
normally distributed	正态分布
null hypothesis	零假设
odds ratio	比数比
ordered logit/probit	次序 logit/probit
ordinal outcomes	有序的结果
ordinal dependent variable	有序因变量
Ordinary Least Squares(OLS)	普通最小二乘法
parallel slopes cumulative model	平行斜率累积模型
parallel slopes	平行斜率
parameter	参数
perfect prediction	理想预测
probability density function	概率密度函数
proportional-odds model	比例比数模型
qualitative choice model	定性选择模型
random utility model	随机效用模型
risk-ratio	风险比
sample statistics	样本统计量
simulation	模拟
skewed ordinal data	偏斜有序数据
standard normal variate(SNV)	标准正态变量
stereotype model	构造型模型
subsample	子样本
threshold values	临界值
unbalanced sample	不平衡样本
utility maximization	效用最大化
Wald statistic	沃尔德统计量
Weibull distribution	威布尔分布函数